Inexact inverse iteration using Galerkin Krylov solvers

Berns-Mller, Jrg

Award date:
2003

Awarding institution:
University of Bath

Inexact Inverse Iteration using Galerkin Krylov solvers

submitted by

Jörg Berns-Müller

for the degree of PhD

of the

University of Bath

2003

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Summary

This thesis is concerned with the convergence and efficiency of inexact inverse iteration applied to the standard symmetric and the generalised unsymmetric eigenvalue problem. Here we mean by inexact inverse iteration that the arising linear systems are solved inexactly using an iterative method. Hence inexact inverse iteration is an inner-outer type algorithm. We provide for the standard symmetric eigenvalue problem and for the generalised unsymmetric eigenvalue problem general convergence results. Both convergence results are general in the sense that they are independent of the linear solver applied and that they are applicable to various implementations of inexact inverse iteration.

For the case when Galerkin Krylov solvers are applied to the linear systems we analyse the efficiency of inexact inverse iteration. This efficiency analysis combines convergence results for inexact inverse iteration and for the Galerkin Krylov solver. Based on our approach of combining these results we obtain a-posteriori upper bounds on the number of inner-iterations per outer-iteration and a-posteriori upper bounds on the total number of inner-iterations, so the sum of inner-iterations over all outer-iterations. These bounds enable us to discuss the efficiency of practical methods, i.e. how fast is the inner-outer iteration type algorithm. Using above mentioned bounds we show that variations of inexact inverse iteration using a shift tending to the desired eigenvalue are most efficient. Additionally we extend the approach recently published by Simoncini and Elden to arbitrary positive definite preconditioners for the standard symmetric eigenvalue problem, and to the generalised eigenvalue problem. In case of the standard symmetric eigenvalue problem we show that this approach is most efficient.

Acknowledgements

There are many people I like to acknowledge, without their support I probably would have not completed my work. Above all I have to thank my supervisor Prof. Alastair Spence for his patience, encouragement and advise he has given me over the past four years.

Thanks goes also to the department for the nice working atmosphere and the financial support for some of my travels to participate at several conferences. In particular I would like to thank the numerical analysis group for fruitful discussions, guidance and a good time.

As in the course of my work many numerical tests were carried out, I am particularly thankful for the efforts by the computer support team in keeping the machines up and running. Many thanks to Sarah for not blocking the machines with her programs.

Credit for my advances in mastering the English language has to go to many people, but in particular to Kathy, Steve and Alastair. All others I thank for their patience with my ignorance of the proper use of English.

Fellow students I like to thank for the many nice evenings out, the tours across and around Bath.

For the tremendous support and encouragement I received by my former supervisor Rüdiger Seydel, by my parents and not least by other relatives I am very pleased.

Contents

Chapter 1

Introduction

In this thesis we discuss the effect of inexact solves on inverse iteration. We consider this with respect to the standard symmetric eigenvalue problem

$$A\mathbf{x} = \lambda\mathbf{x}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric ($A = A^T$) and with respect to the generalized unsymmetric eigenvalue problem

$$A\mathbf{x} = \lambda M\mathbf{x}, \tag{1.2}$$

where A and $M \in \mathbb{R}^{n\times n}$ and M is symmetric positive definite (spd). Our interest is geared towards the case where the matrices A and M are large, say being derived by the FE-method from a partial differential equation. Inverse iteration requires the solution of linear systems of the form

$$(A - \sigma M)\mathbf{y} = M\mathbf{x}, \tag{1.3}$$

with $M = I$ for (1.1). If A and M are large, direct methods become impractical and iterative techniques, usually combined with preconditioning, are used instead. As a result we have outer-iteration, which are the iterations of inexact inverse iteration, and an inner-iteration, being the iterations of the iterative linear solver.

There are two main aspects of inexact inverse iteration studied in this thesis. First we analyse the convergence of inexact inverse iteration and second we analyse its efficiency when a Galerkin-Krylov solver is applied to the linear system. By efficiency we mean the overall performance of inexact inverse iteration as an inner-outer type method.

Inverse Iteration, so using exact solves, is a well known and well studied algorithm, which has for some time now been outperformed by other methods. The idea of inverse iteration goes at least as far back as Wielandt (1944) and became popular due

to the many contributions from Wilkinson (1958, 1962, 1965) and many more. For the special case of the Rayleigh quotient iteration Ostrowski provided convergence results for the standard symmetric eigenvalue problem (Ostrowski (1957)) and for the standard unsymmetric eigenvalue problem (Ostrowski (1959, 1960)). For further historical references on inverse iteration see Ipsen (1996).

Here we think of inexact inverse iteration as a method in its own right for finding an eigenvalue and an eigenvector rather than the standard technique of finding an eigenvector given a very accurate estimate for the eigenvalue. Of course nowadays one would almost certainly use a Lanczos/Arnoldi-type algorithm, perhaps in the shift-invert mode, or use a Jacobi-Davidson-type algorithm to solve (1.1) or respectively (1.2), but we believe that an in-depth understanding of the basic inexact inverse iteration algorithm for a simple eigenvalue is beneficial to the understanding of more advanced techniques. Recent advances by Simoncini and Szyld (2003) have improved the understanding of inexact Arnoldi-type methods and explain the observation made in Bouras and Fraysee (2000). Bouras and Fraysee (2000) observed that in the inexact Arnoldi method the residual constraint for the inner-methods can be relaxed when the outer process advances. However a full answer for how to optimise the inexact Arnoldi method is missing. Later based on our efficiency analysis we will see that the observation from Bouras and Fraysee (2000) for the inexact Arnoldi method is in contrast to the situation for inexact inverse iteration.

Also for the Jacobi-Davidson (Sleijpen and van der Vorst (2000)) method applied to the standard symmetric eigenvalue problem considerable progress has recently been achieved, specially for calculating the smallest eigenvalues of a positive definite matrix. First there is the observation that only a few vectors need to be kept in the trial space. This leads to the development of the currently best algorithm for calculating the smallest eigenvalue of the standard symmetric eigenvalue problem in the positive definite case, namely LOBPCG, see Knyazev (2000). However if the smallest few, say five vectors are of interest, then the Jacobi-Davidson type method published by Notay (2003) seems to perform best. Notay (2003) links the convergence of inexact Jacobi-Davidson methods with inexact inverse iteration and balances the inner method against the outer method. While these two algorithms seem to be reasonably well understood, there remain open questions, specially for the generalized eigenvalue problem and also for interior eigenvalues. Based on our experience we believe that the performance of these two algorithms, inexact Arnoldi and Jacobi-Davidson, can be improved using some of the insights we obtain from the study of inexact inverse iteration.

The understanding of inexact inverse iteration applied to the standard symmetric eigenvalue problem made considerable progress in the recent years. A very early paper on the use of iterative methods to solve (1.3) is Ruhe and Wiberg (1972). Inexact inverse iteration for symmetric matrices was discussed in Smit and Paardekooper (1999)

where a general theory, independent of the details of the solver, was presented along with some new eigenvalue bounds. An important recent paper on inexact inverse iteration is Simoncini and Eldén (2002) where a version of inexact Rayleigh quotient iteration is discussed. Several key ideas are introduced specially with regard to the derivation of the appropriate linear system to be solved when approximate Cholesky preconditioning is applied to the linear system (1.3), and with regard to the determination of an appropriate stopping condition in the inner iteration. We shall discuss these ideas in detail in Chapter 3. For non-symmetric matrices a fixed shift inexact inverse iteration algorithm is discussed in Golub and Ye (2000). The basic idea is to rearrange the update equation such that only a much simpler correction equation needs to be solved. Golub and Ye (2000) provide a convergence theory along with an analysis of the choice of tolerance used in the inner solves. Convergence results for non-symmetric matrices are also given in Lai et al. (1997). Other related work on the use of inexact Rayleigh quotient iteration to compute the smallest eigenvalue of generalized Hermitian eigenvalue problems is discussed in Notay (2003) and Knyazev and Neymeyr (2003). Furthermore the inverse correction method published by Rüde and Schmid (1995) and Zaslavsky (1995) as an improvement of the variation of inexact inverse iteration using a fixed shift is a special case of the method proposed by Golub and Ye (2000).

One of the key aspects in this thesis is the study of inexact inverse iteration as an inner-outer iteration type algorithm. In order to study an inner-outer type technique we need a sufficient understanding of the convergence of both methods. Therefore we have to restrict ourselves to considering only a specific class of methods. Here we consider for the inner method Galerkin Krylov solvers, namely MINRES for the symmetric case and GMRES for the unsymmetric case. However our results are also applicable to CR if the corresponding linear systems are positive definite. The restriction to Galerkin Krylov methods is somehow arbitrary, nevertheless this class of methods is widely used in practice and their convergence is well understood.

As the standard symmetric eigenvalue problem allows us to present the results and techniques in more clarity we analyse in *Chapter 2* the convergence of inexact inverse iteration for this special case. Following Parlett (1980) for exact inverse iteration we use for our analysis a splitting of the current eigenvector approximation into two invariant subspaces, one being the sought eigenspace and the other its orthogonal complement. Our analysis will be independent of the method applied to solve the arising linear systems. In order to achieve this independence we postulate a condition on the residual of the linear solves. This residual condition will later link with the efficiency analysis as MINRES minimises this residual. A different point of view has been taken in Smit and Paardekooper (1999) and Neymeyr (2001b). Both use a backward error analysis type of approach to analyse the convergence of inexact inverse iteration, i.e. they interpret that each iteration is performed exactly starting from a perturbed initial

problem. Their approach is connected to our approach as the residual of the linear system can be interpreted as the perturbation. Besides the advantage of using the residual later in the efficiency analysis we find that our approach is more intuitive. The main convergence result generalizes and combines the results of Smit and Paardekooper (1999) and Golub and Ye (2000). Based on this general convergence result we provide convergence results for three variations of inexact inverse iteration. The first variation uses a fixed shift and decreasing tolerance, a convergence result for such an approach has been proven by Smit and Paardekooper (1999). Second we consider an inexact Rayleigh quotient iteration with a fixed tolerance condition, this case is also proven in Smit and Paardekooper (1999) whose result also allows us to deduce the convergence of the third approach using the Rayleigh quotient as shift and a decreasing tolerance. While the fixed shift gives only linear convergence the Rayleigh quotient iteration with fixed tolerance exhibits locally quadratic convergence and the Rayleigh quotient iteration with decreasing tolerance has locally cubic convergence. At the end of Chapter 2 we will provide numerical results illustrating the convergence of these three methods.

In *Chapter 3* we again consider the standard symmetric eigenvalue problem, but now our main focus will be the efficiency of inexact inverse iteration using MINRES as linear solver. In order to state our efficiency results we need two basic results. In the first result we prove under suitable conditions that the number of outer-iterations decreases when the order of convergence of the outer method is improved. The second result provides an upper bound on the residual in MINRES. To derive this bound we use a standard polynomial approach, however we deflate a few critical eigenvalues including the one we seek using a product of special polynomial and Chebyshev polynomials which is a standard technique in the analysis of polynomial based solvers. This special treatment allows us to project out any contributions in the right-hand side corresponding to critical eigenvalues. This projection will play a key role in determining an efficient method. Further this special treatment for critical eigenvalues allows us to link the convergence of the linear solver with the convergence of the outer method. Now combining inexact inverse iteration and unpreconditioned MINRES we obtain our first efficiency result providing an upper a-posteriori bound on the number of inner-iterations per outer iteration and hence an upper bound on the total number of inner-iterations. The usefulness of this a-posteriori upper bound is that it allows us to provide an analysis of the overall efficiency of the inexact inverse iteration algorithm. Numerical experiments show that this analysis describes well both the behaviour of the inner-iterations and the total number of inner-iterations needed to achieve a desired accuracy for the eigenvector approximation. This a-posteriori upper bound is a vital part in determining efficient methods. As a result we observe that methods using a shift tending as quickly as possible to the sought eigenvalue are most efficient. In practice for the symmetric problem this will probably mean that the shifts should be

chosen to be the Rayleigh quotients. This is consistent with the best strategy when direct solves are used and shows that we need not be concerned that the Krylov solver is applied to a matrix which is becoming more and more singular. The explanation lies in the interplay between the shift tending towards the eigenvalue and the right-hand side tending to the corresponding eigenvector, together with the fact that the Krylov solvers handle reasonably well nearly singular systems with only a small number of critical eigenvalues. Similar ideas were explored in Ruhe and Wiberg (1972) and van der Vorst and Vuik (1993).

The situation changes slightly if we consider inexact inverse iteration applied to (1.1) using preconditioned MINRES. Again we provide a-posteriori bounds on the number of inner-iterations per outer-iteration and the total number of inner-iterations. As in the unpreconditioned case it is the best strategy to reduce the number of outer-iterations which in practice probably means to use the Rayleigh quotient as a shift. However, the cost has now a non-favourable dependency on the error angle of the eigenvalue problem. This is in contrast to the unpreconditioned case and is not just a theoretical problem but is confirmed by numerical experiments. This difference motivates the search for a more sophisticated combination of inexact inverse iteration and preconditioned MINRES. One such approach is the method proposed by Simoncini and Eldén (2002), which itself is based on the observation made by Scott (1981), in case of Cholesky preconditioning that tailoring the right-hand side to the preconditioned solver improves the performance of the linear solver. We extend this method to other preconditioners and prove convergence for this method. Based on our practical experience we suggest using the standard residual stopping condition to stop MINRES rather than the stopping condition suggested by Simoncini and Eldén (2002). Tailoring the right-hand side to the solver is so beneficial that the resulting method is, of all variations of inexact inverse iteration using MINRES known to us, the most efficient. We illustrate various numerical tests to support our theory and also provide some benchmark tests where we compare inexact inverse iteration against the cost of calculating an approximation of the sought eigenpair using LOBPCG and against a linear solve using MINRES.

In *Chapter 4* we consider the generalized unsymmetric eigenvalue problem (1.2). In a similar fashion to Chapter 2 we prove the convergence of inexact inverse iteration applied to the generalized eigenvalue problem. Again this convergence result is general in the sense that it is independent of the linear solver applied to the arising linear system and applicable for various variations of inexact inverse iteration. We will use this result to deduce the convergence of the methods proposed by Rüde and Schmid (1995) and Golub and Ye (2000). Further we deduce the convergence of Rayleigh quotient type methods with fixed or decreasing tolerance and the convergence of a fixed shift method using a decreasing tolerance. Additionally we extend the approach of Simoncini and Eldén (2002) to the generalized eigenvalue problem, however this

method exhibits poor (and only linear) convergence. In the unsymmetric case, only methods combining a Rayleigh quotient type of shift and a decreasing tolerance are of higher order, and so we also consider methods which calculate an approximation to the left eigenvector of (1.2) corresponding to the sought right eigenvector. This allows the use of the generalized Rayleigh quotient which is a higher order approximation of the sought eigenvalue while the standard Rayleigh quotient is only linear in the error angle. We illustrate the convergence behaviour of the considered methods towards the end of Chapter 4.

Chapter 5 is an auxiliary chapter providing the understanding of the convergence of GMRES as needed for the efficiency analysis of inexact inverse iteration in Chapter 6. Using Chebyshev polynomials on the eigenvalues of the shifted linear system matrix $A-\sigma M$ we obtain a convergence result for GMRES. Again we deflate critical eigenvalues and thereby project out any contributions in the right-hand side corresponding to the invariant subspace of these critical eigenvalues. As for the standard symmetric case in Chapter 3 this projection will play a key role in the efficiency analysis in Chapter 6.

In *Chapter 6* we analyse the efficiency of inexact inverse iteration applied to the generalized unsymmetric eigenvalue problem. We restrict to the case where the arising linear systems are solved by 'plain' GMRES, that is GMRES without restarts, deflating or augmenting but with preconditioning. The analysis is similar to the analysis in Chapter 3 where we analysed the efficiency of inexact inverse iteration applied to the standard symmetric eigenvalue problem. For the generalized unsymmetric eigenvalue problem we present two key results. The first result provides an upper-bound on the number of inner-iterations per outer-iteration. This upper bound links nicely with our numerical experience. We observe from our theory and our tests that methods using the standard shifted linear system $(A - \sigma M)\mathbf{y} = M\mathbf{x}$ to update the eigenpair iterates show increasing costs of the linear solves when the outer-method progresses. This first key result also shows that methods using the modified right-hand side as the approach from Simoncini and Eldén (2002) or solve some type of correction equation do not suffer this fate. However, our second key result shows that these latter methods are not efficient due to their poorer, i.e. only linear, outer convergence. Again, methods with minimal number of outer-iterations are most efficient. This means in practice that one should probably use the Rayleigh quotient as a shift and a decreasing residual condition. These theoretical observations are supported by several numerical tests.

Chapter 2

Convergence of Inexact Inverse Iteration for the standard symmetric eigenvalue problem

In this chapter we consider inexact inverse iteration applied to the standard symmetric eigenvalue problem

$$A\mathbf{x} = \lambda\mathbf{x},$$

with A real. By inexact inverse iteration we mean that the linear systems that arise are solved only approximately. We require no further knowledge about the linear solver except that a residual condition is satisfied. Therefore our approach can be viewed as an one level analysis of inexact inverse iteration. This approach leads to a convergence result for inexact inverse iteration independent of the linear solver in use. Based on this general result we will provide convergence results for three practical versions (methods) of inexact inverse iteration. These results are again independent of the linear solver.

The general convergence result and hence the result for the practical methods will be based on an one-step bound. This one-step bound links the approximation quality of an iteration to the approximation quality of the previous iteration. Similar bounds have been proven by Smit and Paardekooper (1999) and Simoncini and Eldén (2002) for the symmetric case and Golub and Ye (2000) for the unsymmetric case. Based on this one-step bound we provide a general convergence result for inexact inverse iteration which unifies the results of Smit and Paardekooper (1999), Golub and Ye (2000) and Simoncini and Eldén (2002). Additionally our convergence result provides the cubic convergence of an inexact Rayleigh quotient iteration with decreasing residual constraint.

The chapter starts with a brief discussion of inverse iteration when exact solves are used. We collect a few basic results which are later used to compare with the

Algorithm 1: inverse iteration

Given $\mathbf{x}^0$ with $\|\mathbf{x}^0\|_2 = 1$,
For $i = 0, 1, 2, \ldots$

- Choose σ^i,
- Solve $(A - \sigma^i I)\mathbf{y}^i = \mathbf{x}^i$,
- Update $\mathbf{x}^{i+1} = \mathbf{y}^i / \|\mathbf{y}^i\|_2$,
- Test for convergence

results of inexact inverse iteration. In Section 2.2 we discuss the convergence of inexact inverse iteration. Starting with an one-step bound, derived in Section 2.2.1, we state and prove the general convergence result in Section 2.2.2. The convergence and the rate of convergence of three practical methods is stated and proven in Section 2.2.3. Further in Section 2.2.4 we comment on literature related to the convergence analysis. Section 2.3 contains numerical examples illustrating our convergence results. These will be summarised in Section 2.3.3.

2.1 Inverse iteration - exact solves

Assume the symmetric matrix $A \in \mathbb{R}^{n \times n}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ with corresponding orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Throughout this chapter we consider the eigenvalue problem

$$A\mathbf{x} = \lambda\mathbf{x} \quad \text{with} \quad \|\mathbf{x}\|_2 = 1. \tag{2.1}$$

In particular, we are interested in the computation of an approximation to one specific simple eigenpair, say $(\lambda_1, \mathbf{v}_1)$. The method we consider here is inverse iteration, as given in Algorithm 1. Given a shift σ^i with $|\lambda_1 - \sigma^i| < |\lambda_j - \sigma^i|$ for all i and all $j > 1$ we assume the following ordering of the eigenvalues

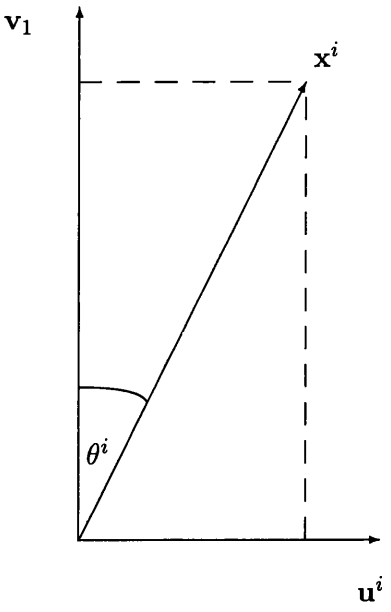
$$|\lambda_1 - \sigma^i| < |\lambda_2 - \sigma^i| \leq \ldots \leq |\lambda_n - \sigma^i| \,. \tag{2.2}$$

For later convenience we assume that $|\lambda_n - \lambda_1| \geq |\lambda_j - \lambda_1|$ for all j and $|\lambda_1 - \sigma| \leq \frac{1}{2}|\lambda_2 - \lambda_1|$. To analyse this algorithm we use a notation similar to Parlett (1980, p. 60).

Assume the orthogonal splitting

$$\mathbf{x}^i = c^i\mathbf{v}_1 + s^i\mathbf{u}^i, \qquad (2.3)$$

with $\|\mathbf{v}_1\|_2 = \|\mathbf{u}^i\|_2 = 1$ and $\mathbf{u}^i \perp \mathbf{v}_1$. If θ^i denotes the angle between $\mathbf{x}^i$ and $\mathbf{v}_1$, so $\theta^i = \angle(\mathbf{v}_1, \mathbf{x}^i)$, then $c^i = \cos\theta^i$ and $s^i = \sin\theta^i$. Thus $\|\mathbf{x}^i - c^i\mathbf{v}_1\|_2 = |s^i|$, so $|s^i|$ is a measure of the convergence of $\mathbf{x}^i$ to $\text{span}\{\mathbf{v}_1\}$. Throughout we use $|s^i|$ or $t^i := \frac{|s^i|}{|c^i|}$ the absolute value of the tangent as a measure of convergence.



Given an eigenvector approximation $\mathbf{x}^i$ we obtain an approximation of the eigenvalue using the Rayleigh quotient (RQ)

$$\begin{aligned}\varrho(\mathbf{x}^i) &:= \frac{(\mathbf{x}^i)^T A\mathbf{x}^i}{(\mathbf{x}^i)^T\mathbf{x}^i} = \lambda_1 + (\mathbf{x}^i)^T(A - \lambda_1 I)\mathbf{x}^i \\ &= \lambda_1 + (c^i\mathbf{v}_1 + s^i\mathbf{u}^i)^T(A - \lambda_1 I)(c^i\mathbf{v}_1 + s^i\mathbf{u}^i) \\ &= \lambda_1 + (s^i)^2\left((\mathbf{u}^i)^T A\mathbf{u}^i - \lambda_1\right), \qquad (2.4)\end{aligned}$$

and therefore $|\lambda_1 - \varrho(\mathbf{x}^i)| \leq |s^i|^2|\lambda_n - \lambda_1|$.

As $|s^i|$ is unknown in practice, we can use the eigenvalue residual as an indicator of its size since the residual is linear in s^i

$$\begin{aligned}\mathbf{r}(\mathbf{x}^i) &:= A\mathbf{x}^i - \varrho(\mathbf{x}^i)\mathbf{x}^i \\ &= s^i(A - \lambda_1 I)\mathbf{u}^i - (s^i)^2\left((\mathbf{u}^i)^T A\mathbf{u}^i - \lambda_1\right)\mathbf{x}^i, \qquad (2.5)\end{aligned}$$

and so $\|\mathbf{r}(\mathbf{x}^i)\|_2 = O(s^i)$. Taking into account that $\|A\mathbf{x}^i - \varrho(\mathbf{x}^i)\mathbf{x}^i\| \leq \|A\mathbf{x}^i - \sigma\mathbf{x}^i\|$ for all σ one obtains

$$\|\mathbf{r}^i\| \leq \|A\mathbf{x}^i - \lambda_1\mathbf{x}^i\| \leq |s^i||\lambda_n - \lambda_1|\,. \qquad (2.6)$$

The rate of convergence for inverse iteration using exact solves is given by

$$\frac{t^{i+1}}{t^i} \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}, \qquad (2.7)$$

see Parlett (1980, p. 62). In Section 2.2 we will derive this inequality (2.7) as a special case of inexact inverse iteration, (2.18). In the case of the Rayleigh quotient iteration

(RQI), that is $\sigma^i = \varrho(\mathbf{x}^i)$, the convergence is given by

$$\frac{t^{i+1}}{(t^i)^3} \leq \frac{|(\mathbf{u}^i)^T A\mathbf{u}^i - \lambda_1|}{|\lambda_2 - \lambda_1| - (s^i)^2 \, |(\mathbf{u}^i)^T A\mathbf{u}^i - \lambda_1|}. \tag{2.8}$$

Ostrowski (1957) was the first to prove the locally cubic convergence. An elegant treatment is given in Parlett (1980, p. 71-74).

Asymptotically $t^{i+1}/(t^i)^3 \to 1$ for the RQI, and $t^{i+1}/t^i \to |\lambda_1 - \sigma^i| \; / \; |\lambda_2 - \sigma^i|$ for any other choice of shift satisfying (2.2). Further the eigenvalue residual has the asymptote $\mathbf{r}(\mathbf{x}^i) \to (\lambda_2 - \lambda_1) s^i \mathbf{v}_2$. These asymptotic results are based on $\mathbf{u}^i \to \mathbf{v}_2$ for $i \to \infty$, which is the case if $|\lambda_2 - \sigma^i| < |\lambda_j - \sigma^i|$ for $j = 3, \ldots, n$ and $\mathbf{v}_2^T \mathbf{u}^0 \neq 0$.

Let $\sigma^i = \lambda_1$, and let a reasonable starting vector $\mathbf{x}^0$ be provided, then (2.7) shows that one iteration is enough to find a perfect approximation to $\mathbf{v}_1$. This has been observed by Wilkinson (1962) who established a $1\frac{1}{2}$ step technique based on a LU decomposition of $A - \sigma^i I$. The $\frac{1}{2}$ step is used to gain a reasonable starting vector, for the 1 iteration.

Obviously a shift closer to the eigenvalue is beneficial to the convergence, so in each iteration one would like to use the best available approximation of λ_1, which leads in practice to different shifts in each iteration. The problem with such an approach is that direct solves, for example using a Gauss-solver, need a factorisation of the shifted linear system. When A is a full matrix the cost of factorising is usually $O(n^3)$ while the costs for one solve are $O(n^2)$ once a factorisation has been performed. So using a new shift in each iteration means a new factorisation is needed. Hence the advantage in the method might be absorbed by the additional cost of factorising. In the inexact case we do not suffer from this additional cost for factorisation and we will show later in Chapter 3, when we discuss which variation of inexact inverse iteration is efficient, that updating the shift in each iteration reduces the overall cost. Another concern with using a shift close to the desired eigenvalue is about the effect of round off errors on the solution. As the shift gets more singular the error in $\mathbf{y}^i$ induced by round off errors might increase dramatically. However as Parlett (1980, pp. 65) shows this error in $\mathbf{y}^i$ might be large but is of no significance to the eigenvalue problem. The main observation is that most of the error in $\mathbf{y}^i$ is in the direction of the sought eigenvector while the error in the other directions remains small.

2.2 Inexact inverse iteration

In contrast to the previous section we consider that the linear systems arising in inverse iteration are only approximately solved, see Algorithm 2.

Further we will use an approach independent of the linear solver and its specifications applied to the linear system. Hence we will assume that the inner method can

Algorithm 2: Inexact inverse iteration

Given $\mathbf{x}^0$,
For $i = 0, 1, 2, \ldots$

- Choose σ^i and τ^i,
- Inexact solve $(A - \sigma^i I)\mathbf{y}^i = \mathbf{x}^i$ such that $\|\mathbf{x}^i - (A - \sigma^i I)\mathbf{y}^i\| \leq \tau^i$,
- Update $\mathbf{x}^{i+1} = \mathbf{y}^i / \|\mathbf{y}^i\|$,
- Test for convergence

find an approximate solution satisfying the tolerance constraint, see Algorithm 2.

We start with a result on the progress being made in one outer iteration in Section 2.2.1. In Section 2.2.2 we prove convergence for inexact inverse iteration in general. Then in Section 2.2.3 we consider three practical versions of inexact inverse iteration, one with a fixed shift, and two versions with the RQ as shift. Finally, in Section 2.2.4 we provide some discussion on related literature.

2.2.1 One Iteration

To analyse the convergence of inexact inverse iteration as given in Algorithm 2 we define the residual as

$$\mathbf{res}^i := \mathbf{x}^i - (A - \sigma^i I)\mathbf{y}^i, \tag{2.9}$$

and extend the previous orthogonal splitting (2.3) to

$$\begin{aligned} \mathbf{x}^i &= c^i \mathbf{v}_1 + s^i \mathbf{u}^i, \\ \mathbf{res}^i &= res_v^i \mathbf{v}_1 + res_u^i \mathbf{u}^i + res_p^i \mathbf{p}^i, \end{aligned} \tag{2.10}$$

with $\mathbf{v}_1$, $\mathbf{u}^i$, and $\mathbf{p}^i$ orthonormal, the second equation defines $\mathbf{p}$ implicitly.

We start our analysis by rewriting the linear solve in Algorithm 2 as an exact equation using the definition of the residual

$$(A - \sigma^i I)\mathbf{y}^i = \mathbf{x}^i - \mathbf{res}^i. \tag{2.11}$$

Further we replace $\mathbf{y}^i$ by $\|\mathbf{y}^i\|_2\ \mathbf{x}^{i+1}$, and use the orthogonal decomposition (2.10)

to obtain

$$\begin{aligned}&\|\mathbf{y}^i\|_2 \left((\lambda_1 - \sigma^i)c^{i+1}\mathbf{v}_1 + s^{i+1}(A - \sigma^i I)\mathbf{u}^{i+1}\right)\\ &= (c^i - res_v^i)\mathbf{v}_1 + (s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i.\end{aligned} \tag{2.12}$$

The orthogonal decomposition splits the invariant subspace spanned by $\mathbf{v}_1$ from the invariant subspace spanned by $\mathbf{v}_2, \ldots \mathbf{v}_n$, so we can split (2.12) into two separate equations, the equation for the cosine part

$$\|\mathbf{y}^i\|_2\ (\lambda_1 - \sigma^i)c^{i+1} = c^i - res_v^i \tag{2.13}$$

and the equation for the sine part

$$\|\mathbf{y}^i\|_2\ s^{i+1}(A - \sigma^i I)\mathbf{u}^{i+1} = (s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i. \tag{2.14}$$

Now we multiply the sine equation (2.14) from the left by $(A - \sigma^i I)^{-1}(I - \mathbf{v}_1\mathbf{v}_1^T)$ and take norms. Due to the orthogonal splitting (2.10), $\|\mathbf{u}^{i+1}\|_2 = 1$ and so the sine equation can be used to derive an upper bound for the absolute value of s^{i+1},

$$\begin{aligned}\|\mathbf{y}^i\|_2|s^{i+1}| &= \|(A - \sigma^i I)^{-1}(I - \mathbf{v}_1\mathbf{v}_1^T)((s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i)\|_2\\ &\leq \|(A - \sigma^i I)^{-1}(I - \mathbf{v}_1\mathbf{v}_1^T)\|_2\ \|(s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i\|_2\\ &\leq \frac{1}{|\lambda_2 - \sigma^i|}\sqrt{(s^i - res_u^i)^2 + (res_p^i)^2}.\end{aligned} \tag{2.15}$$

The same approach can be used to derive a lower bound for $|s^{i+1}|$

$$\begin{aligned}\|\mathbf{y}^i\||s^{i+1}| &= \frac{\|(A - \sigma^i I)\|_2}{\|(A - \sigma^i I)\|_2}\ \|(A - \sigma^i I)^{-1}((s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i)\|_2\\ &\geq \frac{\|(s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i\|_2}{\|(A - \sigma^i I)\|_2}\\ &\geq \frac{1}{|\lambda_n - \sigma^i|}\sqrt{(s^i - res_u^i)^2 + (res_p^i)^2}.\end{aligned} \tag{2.16}$$

By dividing (2.15) and (2.16) by the absolute value of (2.13) we obtain a lower and an upper bound on the tangent

$$\begin{aligned}\frac{|\lambda_1 - \sigma^i|}{|\lambda_n - \sigma^i|}\ \frac{\sqrt{(s^i - res_u^i)^2 + (res_p^i)^2}}{|c^i - res_v^i|}&\\ \leq t^{i+1} \leq&\\ \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}\ \frac{\sqrt{(s^i - res_u^i)^2 + (res_p^i)^2}}{|c^i - res_v^i|}&.\end{aligned} \tag{2.17}$$

In practice, we expect the upper bound to reflect the true value more accurately than the lower one, as for exact solves $\frac{t^{i+1}}{t^i} \to \frac{|\lambda_1-\sigma^i|}{|\lambda_2-\sigma^i|}$, see Section 2.1.

However, in this work we mainly use a simpler form of the upper bound on the tangent (2.17), which uses that $|\mathbf{v}_1^T(\mathbf{x}^i - \mathbf{res}^i)| \geq |c^i| - \|\mathbf{res}^i\|_2$ and

$$\|(I - \mathbf{v}_1\mathbf{v}_1^T)(\mathbf{x}^i - \mathbf{res}^i)\|_2 \;\leq\; |s^i| + \|\mathbf{res}^i\|_2 \,.$$

Together we gain for the new tangent the upper bound which we frequently refer to as the one-step bound

$$t^{i+1} \;\leq\; \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}\,\frac{|s^i| + \|\mathbf{res}^i\|_2}{|c^i| - \|\mathbf{res}^i\|_2}. \tag{2.18}$$

As $\|\mathbf{res}^i\|_2 = 0$ for direct solves, above bound (2.18) can be seen as a generalisation of the bound on the convergence rate when exact solves are used (2.7). Further (2.18) is an upper bound on the right-hand side of (2.17). The benefit of this bound is based entirely on its simplicity and its obvious relation to (exact) inverse iteration.

Bounds similar to (2.18) can be found in Smit and Paardekooper (1999), Golub and Ye (2000) and Simoncini and Eldén (2002).

2.2.2 Convergence theory

Based on the one-step bound (2.18) we state the general convergence result in Theorem 2.1. Later in Section 2.2.3 we state the convergence for three practical methods as corollaries of this result.

Theorem 2.1 *Consider inexact inverse iteration, defined by Algorithm 2, applied to* $A \in \mathbb{R}^{n\times n}$, *A symmetric. Assume* $\exists C_1, C_2, \alpha \in \mathbb{R}^+$ *and* $\beta \in [0,1]$ *and* $C_3 \in [0,1)$ *such that for all* $\mathbf{x}^i = c^i\mathbf{v}_1 + s^i\mathbf{u}^i$ *the shift satisfies*

$$|\lambda_1 - \sigma^i| \;\leq\; \min\{C_1\,|s^i|^\alpha, \frac{1}{2}\,|\lambda_2 - \lambda_1|\}$$

and that the residual satisfies

$$\|\mathbf{res}^i\|_2 \;\leq\; \min\{C_2\,|s^i|^\beta, C_3\,|c^i|\}$$

for $\alpha + \beta \geq 1$. *If the initial approximation* $\mathbf{x}^0 = c^0\mathbf{v}_1 + s^0\mathbf{u}^0$ *is such that,* $c^0 \neq 0$, *and*

$$(t^0)^{\alpha+\beta-1} \;<\; \frac{|\lambda_2 - \lambda_1|}{2C_1(1 + C_2)}(1 - C_3), \tag{2.19}$$

then $t^i \to 0$. *Hence* $\mathbf{x}^i \to \mathbf{v}_1$ *and* $\varrho(\mathbf{x}^i) \to \lambda_1$.
Proof: Define

$$C_4 := |s^0|^{\alpha+\beta-1}\, 2C_1(1+C_2)(1-C_3)^{-1}\, |\lambda_2-\lambda_1|^{-1},$$

then with the condition on the initial guess (2.19) and $|s^0| \leq t^0$ we have $C_4 < 1$. From the one-step bound (2.18) we gain

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1-\sigma^i|}{|\lambda_2-\sigma^i|}\,\frac{|s^i|+\|\mathbf{res}^i\|_2}{|c^i|-\|\mathbf{res}^i\|_2} \\ &\leq \frac{2C_1\,|s^i|^\alpha}{|\lambda_2-\lambda_1|}\,\frac{|s^i|+C_2\,|s^i|^\beta}{|c^i|-C_3\,|c^i|} \\ &\leq t^i\frac{|s^i|^{\alpha+\beta-1}}{|s^0|^{\alpha+\beta-1}}C_4. \end{aligned} \tag{2.20}$$

Now we use induction to prove $|s^{i+1}| \leq |s^i|$ on the basis that $|s^i| \leq |s^0|$ which is satisfied for $i = 0$. For $i > 0$ we get by using (2.20) $t^{i+1}/t^i leC_4 < 1$ and therefore $|s^{i+1}| < |s^i|$ for $i \geq 0$. Further (2.20) leads to

$$t^{i+1} \leq C_4 t^i \leq \ldots \leq (C_4)^{i+1} t^0,$$

and so $t^i \to 0$. With (2.4) we gain

$$\begin{aligned} |\varrho(\mathbf{x}^i)-\lambda_1| &\leq (s^i)^2\,|(\mathbf{u}^i)^T A\mathbf{u}^i-\lambda_1| \\ &\leq (t^i)^2\,|\lambda_n-\lambda_1|, \end{aligned}$$

hence $\varrho(\mathbf{x}^i) \to \lambda_1$. As $\mathbf{v}_1$ and $-\mathbf{v}_1$ are the 'same' eigenvector of A corresponding to λ_1, assume without loss of generality that $c^i \geq const > 0$, then $\|\mathbf{x}^i - c^i\mathbf{v}_1\|_2 \leq |s^i|$ and hence $\mathbf{x}^i \to \mathbf{v}_1$. □

The conditions in Theorem 2.1 are not of direct practical use. In general they state if either the shift converges to λ_1 or the tolerance converges to zero and an adequate initial guess $\mathbf{x}^0$ is given, then $\mathbf{x}^i \to \mathbf{v}_1$ and $\varrho(\mathbf{x}^i) \to \lambda_1$. However this result will play a key role in the efficiency analysis in Chapter 3. Further, condition (2.19) is marginally more restrictive than needed, as $(t^0)^{\alpha+\beta-1}$ could be replaced by $|s^0|^{\alpha+\beta-1}$. Later in Chapter 3 we need condition (2.19) in the form stated here.

We note without proof that $\mathbf{u}^i \not\to \mathbf{v}_2$ unless $|\lambda_2-\sigma^i| < |\lambda_j-\sigma^i|$ for $j = 3,\ldots,n$ and either $\tau^i/\,|s^i| \to 0$ or $\mathbf{res}^i \to \mathrm{span}(\mathbf{v}_1,\mathbf{v}_2)$. To justify this we observe that

$$\begin{aligned} \mathbf{u}^{i+1} &= \frac{1}{s^{i+1}}(\mathbf{x}^{i+1}-c^{i+1}\mathbf{v}_1) \\ &= \frac{1}{s^{i+1}}((A-\sigma^i I)^{-1}(\mathbf{x}^i-res^i)-c^{i+1}\mathbf{v}_1) \end{aligned}$$

$$= \frac{1}{s^{i+1}}(A-\sigma^i I)^{-1}(s^i\mathbf{u}^i - res_u^i\mathbf{u}^i - res_p^i\mathbf{p}^i)$$
$$= \frac{1}{s^{i+1}}(A-\sigma^i I)^{-1}((s^i - res_u^i)\mathbf{u}^i - res_p^i\mathbf{p}^i).$$

In order to achieve $\mathbf{u}^i \to \mathbf{v}_2$ the $\mathbf{p}^i$ component needs to vanish which implies that

$$\frac{res_p^i}{s^i - res_u^i} \to 0. \tag{2.21}$$

If $\tau^i/s^i \to 0$ then (2.21) is satisfied. Now assume $\tau^i/s^i \geq const > 0$, then $s^i - res_u^i \geq const$. Hence $res_p^i/(s^i - res_u^i) \geq res_p^i const$, therefore (2.21) implies $res_p^i \to 0$.

2.2.3 Three practical methods

Next we consider three practical versions for the choice of σ^i and τ^i to be used in inexact inverse iteration, Algorithm 2:

- (inexact) inverse iteration with fixed shift and decreasing tolerance,

$$\sigma^i = \sigma^0 \quad \text{and} \quad \tau^i = \min(\tau^0, C_2\,|s^i|), \tag{2.22}$$

- (inexact) Rayleigh quotient iteration with fixed tolerance,

$$\sigma^i = \varrho(\mathbf{x}^i) \quad \text{and} \quad \tau^i = \tau^0, \tag{2.23}$$

- (inexact) Rayleigh quotient iteration with decreasing tolerance,

$$\sigma^i = \varrho(\mathbf{x}^i) \quad \text{and} \quad \tau^i = \min(\tau^0, C_2\,|s^i|). \tag{2.24}$$

As s^i is usually unknown in practice one might use

$$\begin{aligned} \tau^i &:= \min(\tau^0, \widetilde{C}_2\,\|\mathbf{r}^i\|) \quad \text{or} \\ \tau^i &:= \min(\tau^0, \widetilde{C}_2\,\|\mathbf{r}^i\| \,/\, |\varrho(\mathbf{x}^i)|), \end{aligned} \tag{2.25}$$

motivated by (2.6). Smit and Paardekooper (1999) suggest a different approach for choosing τ^i based on data from two outer iterations

$$\tau^i = \min\left(\tau^0, \frac{(1-\nu q^i)q^i}{(1+q^i)\,|\varrho(\mathbf{x}^i) - \sigma^0|}\,\|\mathbf{r}^i\|\right), \tag{2.26}$$

where $q^i := \|\mathbf{r}^i\| \,/\, \|\mathbf{r}^{i-1}\|$ and $\nu \in [1, \frac{|\lambda_2-\sigma^0|}{|\lambda_1-\sigma^0|}]$. As $q^i \to \frac{|\lambda_1-\sigma^0|}{|\lambda_2-\sigma^0|}\nu$, we observe that τ^i is again linear in $|s^i|$. Generalising the tolerance choice of Smit and Paardekooper (1999) to variable shifts is not straight forward as the denominator then contains $\varrho(\mathbf{x}^i) - \sigma^i$.

Another approach is to choose

$$\tau^i := \zeta \tau^{i-1} \tag{2.27}$$

for a fixed $0 < \zeta < 1$, see Golub and Ye (2000). In practice a fixed shift $\sigma^i = \sigma^0$ together with (2.27) leads to linear convergence with a convergence rate

$$\frac{t^{i+1}}{t^i} \approx \max\left(\frac{|\lambda_1 - \sigma^0|}{|\lambda_2 - \sigma^0|}, \zeta\right). \tag{2.28}$$

Golub and Ye (2000) use this approach for the unsymmetric eigenvalue problem, though, obviously the result holds for symmetric problems. We discuss this approach further in Chapter 4.

The convergence results for these three practical methods can be written as Corollaries of Theorem 2.1. However for the the case of a fixed shift (2.22) we reformulate the statement to make it more readable.

Corollary 2.2 *Apply Algorithm 2 with shift and tolerance chosen using (2.22) to* $A \in \mathbb{R}^{n\times n}$, *symmetric. If* $\tau^0 \leq \frac{1}{2}$ *and* $\mathbf{x}^0$ *such that*

$$C_6 := \frac{C_5}{1 - C_5} \frac{1 + \tau^0}{|c^0| - \tau^0} < 1, \tag{2.29}$$

where $C_5 := |\lambda_1 - \sigma^0| \,/\, |\lambda_2 - \lambda_1|$, *then* $t^i \leq (C_6)^i t^0$, *hence* $\varrho(\mathbf{x}^i) \to \lambda_1$ *and* $\mathbf{x}^i \to \mathbf{v}_1$.

Proof: We use the definition of C_5 and and the definition of inexact inverse iteration using a fixed shift (2.22) to obtain from the one-step bound (2.18)

$$t^{i+1} \leq C_6 t^i \leq \ldots \leq (C_6)^{i+1} t^0. \tag{2.30}$$

By using the same argument as in the proof of Theorem 2.1 we obtain $\varrho(\mathbf{x}^i) \to \lambda_1$ and $\mathbf{x}^i \to \mathbf{v}_1$. □

Similar results have been published by Smit and Paardekooper (1999) and Golub and Ye (2000).

Corollary 2.3 *Apply the RQI with fixed tolerance, that is Algorithm 2 with shift and tolerance chosen using (2.23), to* $A \in \mathbb{R}^{n\times n}$, *symmetric. If* $\tau^0 \leq \frac{1}{2}$ *and* $\mathbf{x}^0$ *such that*

$$t^0 < \frac{1}{5} \frac{|\lambda_2 - \lambda_1|}{|\lambda_n - \lambda_1|} \frac{1}{1 + \tau^0}, \tag{2.31}$$

then $t^i \to 0$ *quadratically.*

Proof: The condition on t^0, (2.31) implies $|s^0| \leq \frac{1}{5}$ and hence $|c^0| \geq \sqrt{1 - \frac{1}{25}} \geq 4/5$.

Now set $C_2 = C_3 = 3/5$, then with (2.23)

$$\tau^i \leq \tau^0 \leq \min\{C_2, C_3\,|c^0|\} \leq \min\{C_2, C_3\,|c^i|\}.$$

Further we use the bound on $|\lambda_1 - \varrho(\mathbf{x}^i)| \leq |s^i|^2|\lambda_n - \lambda_1|$ from Section 2.1 and set $C_1 = |\lambda_n - \lambda_1|$. Next we set $\alpha = 2$ and $\beta = 0$ to obtain

$$\begin{aligned} t^0 &= (t^0)^{\alpha+\beta-1} < \frac{1}{5}\frac{|\lambda_2 - \lambda_1|}{|\lambda_n - \lambda_1|}\frac{1}{1+\tau^0} \\ &\leq \frac{1}{2}\frac{|\lambda_2 - \lambda_1|}{C_1(1+C_2)}(1-C_3)\left(\frac{2}{5}\frac{1}{1-C_3}\right) \\ &\leq \frac{1}{2}\frac{|\lambda_2 - \lambda_1|}{C_1(1+C_2)}(1-C_3). \end{aligned}$$

Now we use the definition of the RQI with fixed tolerance (2.23) together with the fact that the RQ is quadratic in s^i, (2.4) and get

$$\begin{aligned} |\lambda_1 - \sigma^i| &\leq |s^i|^2|\lambda_n - \lambda_1| \\ &\leq \frac{1}{2}|\lambda_2 - \lambda_1|\left(\frac{2}{25}\frac{|\lambda_2 - \lambda_1|}{|\lambda_n - \lambda_1|}\left(\frac{1}{1+\tau^0}\right)^2\right) \\ &< \frac{1}{2}|\lambda_2 - \lambda_1|\,. \end{aligned}$$

Thus the conditions of Theorem 2.1 are satisfied and we get $t^i \to 0$. To prove the quadratic convergence we use the one-step bound (2.18) and the quadratic behaviour of the RQ (2.4). Further we use $|c^i| - \|\mathbf{res}^i\| \geq |c^0| - \tau^i \geq \frac{4}{5} - \frac{1}{2} = \frac{3}{10}$ hence

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}\frac{|s^i| + \|\mathbf{res}^i\|_2}{|c^i| - \|\mathbf{res}^i\|_2} \\ &\leq |s^i|^2\, 2\frac{|\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1|}\frac{1+\tau^0}{\frac{3}{10}} \\ &\leq (t^i)^2\,\frac{20}{3}\frac{|\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1|}(1+\tau^0). \end{aligned}$$

Hence $t^{i+1}/(t^i)^2 \leq const$, so $t^i \to 0$ quadratically. □

The quadratic convergence of the RQI with fixed tolerance is also proven in Smit and Paardekooper (1999). In the following we prove the locally cubic convergence of the RQI with decreasing tolerance.

Corollary 2.4 *Apply the RQI with decreasing tolerance, that is Algorithm 2 with shift and tolerance chosen using (2.24), to $A \in \mathbb{R}^{n\times n}$, symmetric. If the tolerance $\tau^i \leq \tau^0 \leq$*

$\frac{1}{2}$, *and the initial guess* $\mathbf{x}^0 = c^0\mathbf{v}_1 + s^0\mathbf{u}^0$ *is such that*

$$t^0 < \frac{1}{5}\frac{|\lambda_2 - \lambda_1|}{|\lambda_n - \lambda_1|}\frac{1}{1+\tau^0}, \tag{2.32}$$

then $t^i \to 0$ *and locally*

$$t^{i+1}/(t^i)^3 \leq 4\frac{|\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1|}(1+|\lambda_n - \lambda_1|).$$

Proof: The conditions of Corollary 2.3 are satisfied hence $t^i \to 0$. In the limit we have $\|\mathbf{res}^i\| \leq |\lambda_n - \lambda_1||s^i| \leq \frac{1}{4}$, hence

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}\frac{|s^i| + \|\mathbf{res}^i\|_2}{|c^i| - \|\mathbf{res}^i\|_2} \\ &\leq |s^i|^3\ 2\frac{|\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1|}\frac{1+|\lambda_n - \lambda_1|}{|c^i| - |s^i||\lambda_n - \lambda_1|} \\ &\leq (t^{i+1})^3\frac{|\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1|}4(1+|\lambda_n - \lambda_1|) \end{aligned}$$

as $|c^i| \geq \frac{3}{4}$. □

This result is not proven explicitly in Smit and Paardekooper (1999) but is implicit in their convergence result for inexact Rayleigh quotient iteration with fixed tolerance. In Section 2.1 we have seen that the convergence for inverse iteration with exact solves is linear for fixed shifts. The same is true for inexact inverse iteration with fixed shift and decreasing tolerance, see Corollary 2.2. While this is not surprising we have a similar situation for the Rayleigh quotient iteration. Recall, see Section 2.1 or see Ostrowski (1957) or Parlett (1980, Section 4.6), the Rayleigh quotient iteration with exact solves has locally cubic convergence. As proved in Corollary 2.4 the same is true for the inexact Rayleigh quotient iteration with tolerance condition linear in $|s^i|$. However this is not the case when the residual condition τ^i is fixed as in case of Corollary 2.3.

Numerical tests to illustrate the results will be presented in Section 2.3

2.2.4 Related literature

Lai et al. (1997) proposed a version of inexact inverse iteration to find the eigenvalue smallest in magnitude. As their algorithm is designed for a fixed shift and uses a decreasing tolerance, the convergence is only linear. Later in Chapter 3 we show that such linearly converging methods are not competitive, hence we do not discuss them in full detail.

For the symmetric eigenvalue problem Smit and Paardekooper (1999) proved convergence for inexact inverse iteration using a fixed shift and the tolerance update for-

mula (2.26). Further they proved quadratic convergence for the RQI with fixed tolerance. Tests with their stopping condition, which we do not present here, lead to inferior results compared with results based on conditions linear to the residual.

Another version of inexact inverse iteration is the inverse correction method from Rüde and Schmid (1995), we discuss this approach in more detail in Section 3.5. A similar approach has been proposed by Golub and Ye (2000), together with the tolerance update (2.27). Both approaches are designed for a fixed shift and a decreasing tolerance, hence their convergence is only linear. However we do not present these results here.

Neymeyr (2001a,b) considers only the smallest eigenvalue for a positive definite matrix A in his geometrical approach for preconditioned inverse iteration. The analysis is geared towards the use of multigrid as a linear solver.

The preconditioned approach by Simoncini and Eldén (2002) we discuss later in Section 3.6.1 in more detail. However Simoncini and Eldén (2002, Theorem 4.1) derive a result similar to (2.17) for unpreconditioned MINRES solves.

The lower bound on the new tangent, see (2.17), can be used to understand why certain variations of inexact inverse iteration do not converge. In cases reported by Rüde and Schmid (1995) and Hawkins (1999) the residual condition τ^i and the shift σ^i have been fixed, so convergence is not secured. And the lower bound on the new tangent, (2.17) shows that a fixed tolerance τ^i and a fixed shift σ^i lead to stagnation if $\mathbf{res}^i \not\to \mathbf{v}_1$, and $\mathbf{res}^i \not\to \mathbf{0}$. Due to practical experience we know that the residual is unlikely to converge towards $\operatorname{span}\{\mathbf{v}_1\}$, but is not impossible.

Earlier we assumed that the RQ is the best approximation known for the sought eigenvalue. However Ostrowski (1958) states an update on the Rayleigh quotient using two successive iterates of Algorithm 1 with $\sigma^i = \varrho(\mathbf{x}^i)$ which gives a minor advantage over the Rayleigh quotient. The technique from Ostrowski (1958) is based on experimental knowledge in case of exact solves. Such empirical relations break down when the convergence has an undetermined random part, which is the case for inexact inverse iteration. Another approach to improve on the Rayleigh quotient, according to Parlett (1980, p. 149) is the Wilkinson shift, which is basically a 2×2 subspace approximation. However we do not investigate such shifts here, despite their advantage with respect to global convergence.

Even with exact solves RQI lacks global convergence. Therefore in a practical situation when one seeks an eigenvalue in a given interval, it is sensible to start with a fixed shift and switch to the RQI once the RQ is inside the interval. For more on this issue see Szyld (1988). As the conditions in Corollaries 2.3 and 2.4 ensure that the RQ is closer to the eigenvalue of interest than to any other the RQI converges to the desired eigenpair. However this observation is based on the fact that the RQI is more efficient than inverse iteration with a fixed shift. While this is a known fact for direct

solves, in Chapter 3 we shall show that this is also the case for inexact solves.

As we explained in Chapter 1 an extension of the RQI to subspaces is given by Absil et al. (2002). In the one dimensional case the Grassmann RQI is equal to the RQI.

2.3 Numerical Examples

In this section we illustrate the convergence results established in the previous section. To do so we consider two examples 'Poisson' and 'bcsstk09', which is a matrix from `http://gams.nist.gov/MatrixMarket`, which we introduce in Section 2.3.1. The example 'bcsstk09' will be used to illustrate the convergence behaviour in practical situations. Due to the limitation of machine precision the difference between quadratic and cubic convergence is not observable using 'bcsstk09', so we consider the example 'Poisson' using variable precision arithmetic to demonstrate the differences between quadratic and cubic convergence.

We start by defining some abbreviations for the considered practical methods in Section 2.3.1. Then in Section 2.3.2 we consider specific tests for the considered examples and methods. These tests will be discussed in light of the convergence analysis of Section 2.1. Finally in Section 2.3.3 we give a short summary.

2.3.1 Notation and examples

In the following definition of some abbreviations we use the constants $\widetilde{C}_1$, $\widetilde{C}_2$, and $\widetilde{C}_3$ referring to the constants C_1, C_2, and C_3 respectively in Theorem 2.1. The tilde indicates that the use of these constants is altered.

Invit stands for inverse iteration with fixed shift and decreasing tolerance, $\sigma^i = \varrho^0$ and $\tau^i = \min\{\widetilde{C}_2\, |\varrho^i|^{-1}\|r^i\|_2, \widetilde{C}_3\}$.

RQIf is the Rayleigh quotient iteration with fixed tolerance, $\sigma^i = \varrho^i$ and $\tau^i = \widetilde{C}_3$.

RQId is the Rayleigh quotient iteration with decreasing tolerance, $\sigma^i = \varrho^i$ and $\tau^i = \min\{\widetilde{C}_2\, |\varrho^i|^{-1}\|r^i\|_2, \widetilde{C}_3\}$.

This list will be extended in Chapter 3, when we discuss more variations of inexact inverse iteration. Additional to the residual stopping condition $\|\mathbf{res}^i\| \leq \tau^i$ we stop the inner iteration when the target for the outer iteration is reached. This is primarily done to improve the robustness of the methods. Robustness and stopping conditions to improve the robustness are discussed later in Section 3.7.1.

Poisson eigenvalue problem on a rectangular domain with aspect ratio 1/1.3 and Dirichlet boundary conditions. For the discretisation we use thirteen grid points

per direction and a second order central finite difference scheme. We consider only the smallest eigenvalue of this 121×121 matrix,

i^{th} smallest	1	2	121
value	15.6	32.6	901.2

bcsstk09 is a real symmetric matrix from Matrix-Market which we use for this example. All tests presented have the same starting vector $\mathbf{x}^0 = c^0\mathbf{v}_1 + s^0\mathbf{u}^0$ where $t^0 = s^0/c^0 = 0.02$. We try to find the 20^{th} smallest eigenvalue of the 1083×1083 matrix. In the following table we summarise those eigenvalues which describe the difficulty for the corresponding tests.

i^{th} smallest	19	20	21	1083
value	3.7e+5	4.1e+5	4.4e+5	6.7e+7

2.3.2 Results and interpretation

In this section we illustrate the convergence behaviour for the methods defined in Section 2.3.1. First we consider the example 'bcsstk09' and compare the methods derived from the theory as developed in Section 2.2. Then we use the example 'Poisson' to illustrate the difference between the quadratic convergence of RQIf and the cubic convergence of RQId.

In order to compare the test result easily we restrict to examples starting with the same starting vector, which is constructed by multiplying the matrix of eigenvectors by a random vector. Further we restrict ourselves to a few tests in order to illustrate a few key points. However we completed tests with different matrices, with different linear solvers, with different initial conditions and all these tests underline the findings we will present here unless otherwise stated.

Test 2.1 *Consider Invit using unpreconditioned MINRES applied to 'bcsstk09'. We try to find the* 20^{th} *smallest eigenvalue and use* $\|\mathbf{r}^i\| \,/\, |\varrho^i| \leq 10^{-10}$ *as stopping condition for the outer method. We present tests for two different sets of parameter values in Table 2.1. The first set of parameters is* $\widetilde{C}_2 = 1$ *and* $\widetilde{C}_3 = 0.2$ *while the second is* $\widetilde{C}_2 = 0.25$ *and* $\widetilde{C}_3 = 0.02$. *In Table 2.1 we list the norm of the eigenvalue residual* $\|\mathbf{r}^i\|$, *the tangent* t^i *and the progress* t^{i+1}/t^i *against the number of outer iterations* i. *The tangent is calculated using an eigenvector approximation with relative accuracy* 10^{-12}.

For Test 2.1, Corollary 2.2 predicts linear convergence with a convergence rate of at least $t^{i+1}/t^i \leq 0.538$ for exact solves. From the test results, Table 2.1, we see that this rate of convergence can be obtained with inexact solves. We observed this effect only when using unpreconditioned MINRES as a linear solver. Occasionally when using unpreconditioned MINRES we experienced in the early stages of Invit that the convergence rate was significantly better than the convergence bound would suggest.

	$\widetilde{C}_2 = 1$ $\widetilde{C}_3 = 0.2$			$\widetilde{C}_2 = 0.25$ $\widetilde{C}_3 = 0.02$		
i	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$
0	5.0e+05	2.0e-02	2.07e-01	5.0e+05	2.0e-02	1.05e-01
1	2.7e+03	4.1e-03	3.43e-01	3.8e+02	2.1e-03	2.68e-01
2	1.0e+02	1.4e-03	3.62e-01	3.0e+01	5.6e-04	4.47e-01
3	2.1e+01	5.1e-04	4.61e-01	1.0e+01	2.5e-04	5.06e-01
4	9.4e+00	2.4e-04	5.01e-01	5.0e+00	1.3e-04	5.24e-01
5	4.7e+00	1.2e-04	5.15e-01	2.6e+00	6.7e-05	5.30e-01
6	2.4e+00	6.1e-05	5.20e-01	1.4e+00	3.5e-05	5.32e-01
7	1.3e+00	3.2e-05	5.22e-01	7.4e-01	1.9e-05	5.34e-01
8	6.6e-01	1.7e-05	5.24e-01	3.9e-01	1.0e-05	5.35e-01
9	3.4e-01	8.7e-06	5.25e-01	2.1e-01	5.4e-06	5.35e-01
10	1.8e-01	4.6e-06	5.25e-01	1.1e-01	2.9e-06	5.36e-01
11	9.5e-02	2.4e-06	5.26e-01	6.0e-02	1.5e-06	5.36e-01
12	5.0e-02	1.3e-06	5.25e-01	3.2e-02	8.3e-07	5.37e-01
13	2.6e-02	6.6e-07	5.25e-01	1.7e-02	4.4e-07	5.37e-01
14	1.4e-02	3.5e-07	5.26e-01	9.3e-03	2.4e-07	5.37e-01
15	7.2e-03	1.8e-07	5.26e-01	5.0e-03	1.3e-07	5.37e-01
16	3.8e-03	9.6e-08	5.26e-01	2.7e-03	6.9e-08	5.37e-01
17	2.0e-03	5.1e-08	5.26e-01	1.4e-03	3.7e-08	5.37e-01
18	1.1e-03	2.7e-08	5.26e-01	7.7e-04	2.0e-08	5.37e-01
19	5.5e-04	1.4e-08	5.27e-01	4.1e-04	1.1e-08	5.37e-01
20	2.9e-04	7.4e-09	5.27e-01	2.2e-04	5.7e-09	5.37e-01
21	1.5e-04	3.9e-09	5.27e-01	1.2e-04	3.1e-09	5.37e-01
22	8.1e-05	2.1e-09	1.07e-01	6.4e-05	1.6e-09	1.29e-01
23	4.0e-05	2.21e-10		3.0e-05	2.12e-10	

Table 2.1: Invit using MINRES on 'bcsstk09' (Test 2.1)

	RQIf $\widetilde{C}_3 = 0.2$			RQId $\widetilde{C}_2 = 1$ and $\widetilde{C}_3 = 0.2$		
	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$
0	5.0e+05	2.0e-02	2.07e-01	5.0e+05	2.0e-02	2.07e-01
1	2.7e+03	4.1e-03	8.05e-04	2.7e+03	4.1e-03	2.55e-05
2	4.2e-01	3.3e-06	2.89e-05	1.6e-02	1.1e-07	3.88e-03
3	3.9e-05	9.62e-11		3.8e-05	4.10e-10	

Table 2.2: RQIf and RQId using MINRES on 'bcsstk09' (Test 2.2)

However this effect depends on the initial approximation $\mathbf{x}^0$ and is linked with the regularization effect of MINRES, see Kilmer and Stewart (1999). Considerable gains can be obtained in the first outer iteration when the starting vector is biased towards the eigenvectors corresponding to eigenvalues furthest from the sought eigenvalue.

Test 2.2 *We repeat Test 2.1 for RQIf and RQId. The corresponding results are presented in Table 2.2. We only show data for one set of parameters each,* $\widetilde{C}_3 = 0.2$ *for RQIf while* $\widetilde{C}_2 = 1$ *and* $\widetilde{C}_3 = 0.2$ *for RQId.*

While Corollary 2.3 predicts only quadratic convergence for RQIf, Corollary 2.4 predicts cubic convergence for RQId. However in the corresponding test they hardly differ. Both algorithms start with the same shift σ^0 and tolerance τ^0, hence they obtain the same tangent t^1 and the same RQ $\sigma^1 = \varrho^1$. Then in the next iteration the tolerance differs marginally, hence we expect a slightly better improvement for RQId than for RQIf in iteration $i = 2$. However the improvement is considerably better, RQId is about a factor 30 better than RQIf, the latter exhibiting quadratic convergence. This additional improvement is due to the erratic directions in the residual and can have positive and negative results. More importantly these changes do not violate the one step bound (2.18).

Test 2.3 *Here we repeat Tests 2.1 and 2.2 using preconditioned MINRES. The preconditioner is constructed using the MatLab routine* cholinc *with droptol* $= 10^{-2}$. *Corresponding results are in Tables 2.3 and 2.4.*

While for Invit using unpreconditioned MINRES we observed a convergence rate better than the convergence bound for exact inverse iteration we now recover the convergence ratio $t^{i+1}/t^i \approx 0.538$. Comparing Tables 2.1 and 2.3 as well as 2.2 and 2.4 we see that the change of the solver has no significant effect on the outer convergence.

Due to the limitations of the machine precision the results are inconclusive with regard to the actual convergence behaviour of RQIf and RQId. Hence we consider a test using variable precision arithmetic.

Test 2.4 *We use RQIf and RQId on 'Poisson' together with unpreconditioned MINRES. In all test runs we try to find the smallest eigenvalue to an accuracy of* $t^N \approx 10^{-80}$. *Therefore we use as a stopping condition for the outer method* $\|\mathbf{r}^i\|_2 \,/\, |\varrho^i|$. *The results are presented in Table 2.5.*

As the test matrix is only of size 121×121 the next iteration using unpreconditioned MINRES would use 121 inner iteration k^i and would therefore be exact. Further the comparison solution needed to calculate t^i only has 128 decimal digits accuracy. Hence we used the stopping condition $t^i \leq 10^{-80}$. The results presented in Table 2.5 reflect the theoretical prediction. We additionally observe that the quadratically converging RQIf only needs one outer iteration more than the cubicly converging RQId. We also tabulate the number of inner iterations k^i, we will discuss these in more detail

	$\widetilde{C}_2 = 1$ $\widetilde{C}_3 = 0.2$			$\widetilde{C}_2 = 0.25$ $\widetilde{C}_3 = 0.02$		
i	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$
0	5.0e+05	2.0e-02	1.04e+00	5.0e+05	2.0e-02	6.87e-02
1	1.6e+03	2.1e-02	5.06e-01	3.9e+02	1.4e-03	4.67e-01
2	4.2e+02	1.1e-02	5.25e-01	2.6e+01	6.4e-04	5.24e-01
3	2.2e+02	5.5e-03	5.30e-01	1.3e+01	3.4e-04	5.33e-01
4	1.2e+02	2.9e-03	5.33e-01	7.0e+00	1.8e-04	5.36e-01
5	6.1e+01	1.6e-03	5.34e-01	3.7e+00	9.6e-05	5.37e-01
6	3.3e+01	8.3e-04	5.35e-01	2.0e+00	5.2e-05	5.37e-01
7	1.7e+01	4.5e-04	5.36e-01	1.1e+00	2.8e-05	5.37e-01
8	9.3e+00	2.4e-04	5.36e-01	5.8e-01	1.5e-05	5.37e-01
9	5.0e+00	1.3e-04	5.37e-01	3.1e-01	8.0e-06	5.37e-01
10	2.7e+00	6.9e-05	5.37e-01	1.7e-01	4.3e-06	5.38e-01
11	1.4e+00	3.7e-05	5.37e-01	9.0e-02	2.3e-06	5.38e-01
12	7.7e-01	2.0e-05	5.37e-01	4.8e-02	1.2e-06	5.38e-01
13	4.2e-01	1.1e-05	5.38e-01	2.6e-02	6.7e-07	5.38e-01
14	2.2e-01	5.7e-06	5.38e-01	1.4e-02	3.6e-07	5.38e-01
15	1.2e-01	3.1e-06	5.38e-01	7.5e-03	1.9e-07	5.38e-01
16	6.5e-02	1.7e-06	5.38e-01	4.0e-03	1.0e-07	5.38e-01
17	3.5e-02	8.9e-07	5.38e-01	2.2e-03	5.6e-08	5.38e-01
18	1.9e-02	4.8e-07	5.38e-01	1.2e-03	3.0e-08	5.38e-01
19	1.0e-02	2.6e-07	5.38e-01	6.3e-04	1.6e-08	5.38e-01
20	5.4e-03	1.4e-07	5.38e-01	3.4e-04	8.7e-09	5.38e-01
21	2.9e-03	7.4e-08	5.38e-01	1.8e-04	4.7e-09	5.38e-01
22	1.6e-03	4.0e-08	5.37e-01	9.8e-05	2.5e-09	5.38e-01
23	8.4e-04	2.2e-08	5.38e-01	5.3e-05	1.3e-09	5.40e-01
24	4.5e-04	1.2e-08	5.38e-01	3.8e-05	7.3e-10	
25	2.4e-04	6.2e-09	5.38e-01			
26	1.3e-04	3.3e-09	5.38e-01			
27	7.0e-05	1.8e-09	5.38e-01			
28	4.0e-05	9.69e-10				

Table 2.3: Invit using prec. MINRES on 'bcsstk09' (Test 2.3)

	RQIf $\widetilde{C}_3 = 0.2$			RQId $\widetilde{C}_2 = 1$ and $\widetilde{C}_3 = 0.2$		
	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$	$\| r \|$	t^i	$\frac{t^{i+1}}{t^i}$
0	5.0e+05	2.0e-02	1.04e+00	5.0e+05	2.0e-02	1.04e+00
1	1.6e+03	2.1e-02	7.67e-04	1.6e+03	2.1e-02	4.11e-04
2	2.7e+00	1.6e-05	4.82e-06	3.4e-01	8.6e-06	1.15e-05
3	1.9e-05	7.70e-11		2.5e-05	9.82e-11	

Table 2.4: RQIf and RQId using prec. MINRES on 'bcsstk09' (Test 2.3)

	RQIf		RQId	
i	$\log_{10} t^i$	k^i	$\log_{10} t^i$	k^i
0	-0.12	19	-0.14	15
1	-1.41	19	-1.62	24
2	-3.85	33	-4.33	45
3	-9.03	50	-12.90	78
4	-19.46	76	-36.19	113
5	-40.72	108	-82.66	
6	-82.96			
		305		275

Table 2.5: Cubic is better than quadratic (Test 2.4)

in Chapter 3. However we observe that the number of inner iterations per outer iteration is larger when using RQId than when using RQIf, though the overall number of MINRES iteration needed for RQId is 30 fewer than for RQIf. The understanding of these effects will be one aim of Chapter 3.

2.3.3 Conclusion

In Section 2.2 we proved the convergence of inexact inverse iteration for general choices of σ^i and τ^i independent of the linear solver applied. We then specified three practical versions of inexact inverse iteration, which we later called Invit, RQIf and RQId. Our convergence analysis for Invit and RQIf extents the results from Golub and Ye (2000) and Smit and Paardekooper (1999). We also showed how the cubic convergence behaviour of exact inverse iteration can be recovered.

Next we provided some numerical examples to illustrate our convergence results. These numerical results confirm our theoretical findings. Further the numerical results indicate that the difference between quadratic and cubic convergence is negligible for limited machine precision. So it is not worth the effort to obtain cubic convergence. Finally we observed that the cubically converging RQId uses fewer MINRES iterations overall, despite this it uses more MINRES iterations per outer iteration than RQIf.

So far we only have an understanding of the outer convergence but no insight in the interplay between the outer and inner iterations. In order to study this interplay we have to specify a linear solver. This will be done in Chapter 3, where we also analyse some other variations of inexact inverse iteration.

Chapter 3

Efficient Variations of Inexact Inverse Iteration using MINRES

In Chapter 1 we explained the need for finding an efficient variation of inexact inverse iteration. Such a variation takes its strength from the interplay with the linear solver in use, meaning a certain variation might be optimal if MINRES is used as linear solver, but the same variation might perform poorly when for example MultiGrid is used for the linear solve. Hence we will focus on two particular solvers which will be MINRES with and without preconditioning. The decision to use MINRES is based on a variety of reasons including the amount of storage required by the linear solver. Further, the theory for MINRES links directly with the outer convergence theory as MINRES minimises the residual norm over a (Krylov) subspace which is iteratively extended. In Chapter 2 we proved convergence of inexact inverse iteration for general choices of shifts σ^i and residual tolerance constraints τ^i. While the convergence results only give insight which choices for σ^i and τ^i are preferable for the outer iteration, they do not give much help in understanding the overall performance. One task of this chapter is to analyse the effect the choice of σ^i and τ^i has on the performance of MINRES with respect to the eigenvalue problem. Based on this analysis, which we call the 'efficiency analysis', we observe how σ^i and τ^i need to be chosen to gain an efficient method. A key result of the efficiency analysis is that methods with shifts converging towards the sought eigenvalue are most efficient. Further by studying some variations of inexact inverse iteration as suggested by Simoncini and Eldén (2002), we observe that a right hand side tailored to the preconditioner used in MINRES reduces the cost of a linear solve.

In case the shift converges to the sought eigenvalue the linear systems get harder to solve. The application of GMRES to such systems was studied in Brown and Walker (1997). As MINRES is a special implementation of GMRES for symmetric matrices, the results of Brown and Walker (1997) are also valid here. An almost singular system

can cause problems especially if a good solution of the system is needed, that is the error in the solution should be small. However we are not primarily interested in solving the linear system but gaining a good approximation of the sought eigenvector. As the analysis of van der Vorst and Vuik (1993) shows, the convergence of a Krylov technique like MINRES is not hampered too much by a few critical eigenvalues. We will explore this in more detail in the convergence analysis for MINRES. Nevertheless a possible danger of unreliability remains as the linear system is inconsistent. Due to the round off errors, the system will hardly be ever exactly singular, but almost singular. However to maintain robustness of MINRES with respect to the eigenvalue problem we will discuss some additional stopping conditions including monitoring the eigenvalue inside of MINRES.

The scope of this Chapter is as follows. First in Section 3.1 we discuss independent of any particular linear solver how the number of outer iterations is effected by the choice of σ^i and τ^i. Then in Section 3.2 we discuss the convergence of MINRES and provide a bound on the linear solve residual tailored to our later needs. Then in Section 3.3 we link the convergence results of inexact inverse iteration to the cost when MINRES is used to solve the linear systems. By these means a bound on the overall cost of solving the eigenvalue problem is obtained. This result is then used to show that using a shift converging towards the sought eigenvalue reduces the overall cost. We extend the analysis of Section 3.3 for the use of preconditioned MINRES in Section 3.4. In Section 3.5, we consider the Inverse Correction Method suggested by Rüde and Schmid (1995), as an alternative to inexact inverse iteration. We will extend our convergence and efficiency analysis to this case. In Section 3.6 we analyse a generalisation of inexact inverse iteration using modified right-hand sides. This method is motivated by Scott (1981) and extends the ideas of Simoncini and Eldén (2002) to arbitrary but fixed positive definite preconditioners. Convergence and efficiency results as in previous sections are given. The discussion on stopping conditions and robustness is presented in Section 3.7. Section 3.8 is devoted to numerical experiments illustrating the quality of the theoretical results. In this test section we also compare the different variations of inexact inverse iteration, and observe that our newly proposed method outperforms other versions of inexact inverse iteration with respect to reliability and efficiency.

3.1 Number of outer iterations

In this section we show how the number of outer iterations, that is the number of iterations of inexact inverse iteration, $\mathcal{N}$, depends on the shift, the tolerance parameters and the gap $|\lambda_2 - \lambda_1|$. Therefore we derive an upper bound on $\mathcal{N}$ and prove the (almost obvious) result that this bound decreases when the outer convergence rate increases.

We note that the analysis we use in this section is independent of the linear solver applied to the linear system.

Definition 3.1 *Given sequence* (σ^i) *and* (τ^i) *and a constant* $\gamma > 0$, *then define* $\mathcal{N}$ *to be the number of outer iterations needed to improve a current eigenvector approximation* $\mathbf{x}^0$ *with the approximation quality* t^0 *by a factor* $10^{-\gamma}$, *where* $\gamma > 0$.

For example a method to choose σ^i and τ^i is given by the RQI with fixed tolerance (i.e. $\sigma^i = \varrho(\mathbf{x}^i)$ and $\tau^i = \tau^0$), or in a more abstract fashion for example by σ^i such that $|\lambda_1 - \sigma^i| \leq |s^i|^\alpha$ and $\tau^i = |s^i|^\beta$

Obviously $\mathcal{N}$ depends on γ, the shifts σ^i, and the tolerance constraints τ^i as chosen in each iteration. As the following result shows $\mathcal{N}$ also depends on the starting value as t^0 is fixed given an initial approximation $\mathbf{x}^0$.

Lemma 3.2 *Consider inexact inverse iteration, defined by Algorithm 2, for symmetric* $A \in \mathbb{R}^{n\times n}$, *and assume the conditions of Theorem 2.1 hold. Further assume that* $\mathcal{N}$ *is such that* $t^{\mathcal{N}} \leq 10^{-\gamma} t^0$. *Then for* $\alpha + \beta > 1$, $\mathcal{N} \leq 1 + [\mathcal{N}^*]$, *where for* $\alpha + \beta > 1$,

$$\mathcal{N}^* := \frac{1}{\log(\alpha+\beta)} \log \left(\frac{\log \frac{1}{t^0 10^{-\gamma}} + \frac{1}{\alpha+\beta-1} \log \frac{1}{C_4}}{\log \frac{1}{t^0} + \frac{1}{\alpha+\beta-1} \log \frac{1}{C_4}} \right), \tag{3.1}$$

and for $\alpha + \beta = 1$, $\mathcal{N} \leq 1 + [\mathcal{N}^*]$, *where*

$$\mathcal{N}^* := \frac{\log 10^{\gamma}}{|\log C_4|}. \tag{3.2}$$

Here $C_4 := 2C_1 \, |\lambda_2 - \lambda_1|^{-1} \, (1 + C_2)(1 - C_3)^{-1}$.

Proof: To simplify the notation we define $\delta := \alpha + \beta$.

As the conditions of Theorem 2.1 hold, $t^i \to 0$, and for any $\gamma > 0$ there exists $\mathcal{N}$ such that $t^{\mathcal{N}} \leq t^0 10^{-\gamma} < t^{\mathcal{N}-1}$.

Next we use the one-step bound, (2.18) to obtain

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \frac{|s^i| + \|\mathbf{res}^i\|_2}{|c^i| - \|\mathbf{res}^i\|_2} \\ &\leq |s^i|^\alpha \, \frac{2C_1}{|\lambda_2 - \lambda_1|} \, \frac{|s^i|^\beta}{|c^i|} \, \frac{1 + C_2}{1 - C_3} \\ &= t^i \, |s^i|^{\delta - 1} \, C_4. \end{aligned}$$

Using this argument repeatedly we observe

$$\begin{aligned} t^0 10^{-\gamma} < t^{\mathcal{N}-1} &\leq C_4 \left(t^{\mathcal{N}-2}\right)^\delta \leq C_4 \left(C_4 \left(t^{\mathcal{N}-3}\right)^\delta \right)^\delta \\ &\leq \ldots \leq (C_4)^{1 + \delta + \delta^2 + \ldots + \delta^{\mathcal{N}-2}} \left(t^0\right)^{\delta^{\mathcal{N}-1}}. \end{aligned} \tag{3.3}$$

For $\delta = 1$, we obtain $\mathcal{N} \leq 1 + \frac{\gamma \log 10}{|\log C_4|}$, and as $\mathcal{N} \in \mathbb{N}$ we have proven the statement for $\alpha + \beta = 1$.

For the case $\delta > 1$, we obtain from (3.3)

$$t^0 10^{-\gamma} \leq (C_4)^{\frac{\delta^{\mathcal{N}-1}-1}{\delta-1}} (t^0)^{\delta^{\mathcal{N}-1}},$$

which gives by taking logarithms

$$\log(t^0 10^{-\gamma}) \leq \frac{\delta^{\mathcal{N}-1}-1}{\delta-1} \log C_4 + \delta^{\mathcal{N}-1} \log t^0.$$

Rearranging yields

$$\log(t^0 10^{-\gamma}) + \frac{1}{\delta-1} \log C_4 \leq \delta^{\mathcal{N}-1} \left(\log t^0 + \frac{1}{\delta-1} \log C_4 \right).$$

The conditions ensure that $(t^0)^{\delta-1} C_4 < 1$, hence $\log t^0 + (\delta-1)^{-1} \log C_4 < 0$. Hence, dividing by the second factor on the right-hand side leads to

$$\delta^{\mathcal{N}-1} \leq \frac{\log \frac{1}{t^0 10^{-\gamma}} + \frac{1}{\delta-1} \log \frac{1}{C_4}}{\log \frac{1}{t^0} + \frac{1}{\delta-1} \log \frac{1}{C_4}}. \tag{3.4}$$

Again we take the logarithm, divide by $\log \delta$, and substitute back in $\delta = \alpha + \beta$, to obtain

$$\mathcal{N} \leq 1 + \frac{1}{\log(\alpha+\beta)} \log \left(\frac{\log \frac{1}{t^0 10^{-\gamma}} + \frac{1}{\alpha+\beta-1} \log \frac{1}{C_4}}{\log \frac{1}{t^0} + \frac{1}{\alpha+\beta-1} \log \frac{1}{C_4}} \right).$$

We conclude the proof by noting that $\mathcal{N} \in \mathbb{N}$, hence $\mathcal{N} \leq 1 + [\mathcal{N}^*]$. □

We remark that in Lemma 3.2 we split the cases $\alpha + \beta = 1$ and $\alpha + \beta > 1$. However this split does not lead to a discontinuity since for $\alpha + \beta = 1$ we have $C_4 < 1$ and one verifies by using the rule of de l'Hospital that

$$\lim_{\delta \to 1} \left(\frac{1}{\log \delta} \log \frac{\log \frac{1}{t^0 10^{-\gamma}} + \frac{1}{\delta-1} \log \frac{1}{C_4}}{\log \frac{1}{t^0} + \frac{1}{\delta-1} \log \frac{1}{C_4}} \right) = \frac{\log 10^{\gamma}}{|\log C_4|}.$$

In practice it is well known that, for example, the RQI needs fewer iterations than inverse iteration using a fixed shift. However for mathematical rigour we now state and prove a more general version of this practical experience.

Lemma 3.3 *Under the assumptions of Lemma 3.2, $\mathcal{N}^*$, as defined in Lemma 3.2, is decreasing in $\alpha + \beta$.*

Proof: Define δ and C_4 according to Lemma 3.2. Additionally we define

$$\begin{aligned} g(\delta) &:= \frac{\log \frac{1}{t^0 10^{-\gamma}} + \frac{1}{\delta - 1} \log \frac{1}{C_4}}{\log \frac{1}{t^0} + \frac{1}{\delta - 1} \log \frac{1}{C_4}} \\ &= 1 + \frac{\gamma \log 10}{\log \frac{1}{t^0} + \frac{1}{\delta - 1} \log \frac{1}{C_4}} \\ \text{and} \quad f(\delta) &:= \frac{1}{\log \delta} \log(g(\delta)). \end{aligned}$$

To show $\mathcal{N}$ is decreasing in δ for $\delta > 1$, we will show that $f'(\delta) < 0$.

As $C_4 (t^0)^{\delta - 1} < 1$ we have

$$\log \frac{1}{t^0} + \frac{1}{\delta - 1} \log \frac{1}{C_4} \quad > \quad 0.$$

Hence $g(\delta) > 1$, and $f(\delta) > 0$. Next we note

$$g'(\delta) \quad = \quad \frac{(-\gamma \log 10) \left(- \log \frac{1}{C_4} \right)}{\left(\log \frac{1}{t^0} + \frac{1}{\delta - 1} \log \frac{1}{C_4} \right)^2 (\delta - 1)^2},$$

and so $\operatorname{sign}(g'(\delta)) = \operatorname{sign}(\log(1/C_4))$, or equivalently

$$g'(\delta) \leq 0 \quad \Leftrightarrow \quad C_4 \geq 1. \tag{3.5}$$

We draw our attention back to $f(\delta)$ and observe for the case $C_4 \geq 1$

$$f'(\delta) \quad = \quad \frac{-1}{(\log \delta)^2} \frac{1}{\delta} \log(g(\delta)) + \frac{1}{\log \delta} \frac{1}{g(\delta)} g'(\delta) \quad < \quad 0.$$

For $C_4 < 1$ the above approach is indecisive and not of much use. Instead we will study

$$\operatorname{sign}\left(f'(\delta)\right) \quad = \quad \operatorname{sign}\left(\lim_{\varepsilon \to 0} \frac{f(\delta + \varepsilon) - f(\delta)}{\varepsilon} \right).$$

Let $\delta + \varepsilon > 1$ and define $h(\varepsilon) := \log_{\delta + \varepsilon} \delta = \log \delta / \log(\delta + \varepsilon)$, then

$$h'(\varepsilon) \quad = \quad \frac{-\log \delta}{(\log(\delta + \varepsilon))^2} \frac{1}{\delta + \varepsilon} \quad < \quad 0. \tag{3.6}$$

Finally we observe

$$\begin{aligned}
& f(\delta+\varepsilon) - f(\delta) < 0 \\
\Leftrightarrow \quad & \log(\delta) f(\delta+\varepsilon) < \log(\delta) f(\delta) \\
\Leftrightarrow \quad & \exp\left(\log(\delta) f(\delta+\varepsilon)\right) < \exp\left(\log(\delta) f(\delta)\right) \\
\Leftrightarrow \quad & \exp\left(h(\varepsilon)\log(g(\delta+\varepsilon))\right) < \exp\left(h(0)\log(g(\delta))\right) \\
\Leftrightarrow \quad & \left(g(\delta+\varepsilon)\right)^{h(\varepsilon)} < \left(g(\delta)\right)^{h(0)} \\
\Leftarrow \quad & \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(g(\delta+\varepsilon)\right)^{h(\varepsilon)} = h'(\varepsilon)\log(g(\delta+\varepsilon))\left(g(\delta+\varepsilon)\right)^{h(\varepsilon)} g'(\delta+\varepsilon) < 0,
\end{aligned}$$

where the last line is satisfied for $C_4 < 1$ as $h'(\varepsilon) < 0$, see (3.6), and $g(\delta+\varepsilon) > 1$, while $g'(\delta+\varepsilon) > 0$, see (3.5). □

Lemma 3.3 only states the behaviour of $\mathcal{N}^*$ with respect to $\alpha+\beta$ but not with respect to the remaining parameters. To analyse the dependence of $\mathcal{N}^*$ on the remaining parameters we simplify the expression for $\mathcal{N}^*$,(3.1), to

$$\mathcal{N}^* = \frac{1}{\log(\alpha+\beta)} \log\left(1 + \frac{\gamma \log 10}{\log\frac{1}{t^0} + \frac{1}{\alpha+\beta-1}\log\frac{1}{C_4}}\right). \tag{3.7}$$

Differentiating $\mathcal{N}^*$ with respect to t^0 gives for $\alpha+\beta > 1$

$$\begin{aligned}
& \log(\alpha+\beta)\frac{\partial \mathcal{N}^*}{\partial t^0} \\
= \quad & \frac{1}{1 + \dfrac{\gamma\log 10}{\log\frac{1}{t^0} + \frac{1}{\alpha+\beta-1}\log\frac{1}{C_4}}} \frac{\gamma \log 10}{\left(\log\frac{1}{t^0} + \frac{1}{\alpha+\beta-1}\log\frac{1}{C_4}\right)^2} t^0,
\end{aligned} \tag{3.8}$$

which is positive. Similar the derivative with respect to C_4 is also positive. So by tightening the convergence condition or the condition on the initial approximation we decrease C_4 or t^0, hence the bound on the number of outer iterations $\mathcal{N}$ decreases.

3.2 MINRES

3.2.1 Standard convergence analysis

Introduced by Paige and Saunders (1975), MINRES is a Galerkin Krylov technique for solving linear systems. Given $B \in \mathbb{R}^{n\times n}$ nonsingular and $\mathbf{b} \in \mathbb{R}^n$, then MINRES calculates iteratively an approximation $\mathbf{y}_k$ to the solution $\mathbf{y}$ of the linear system $B\mathbf{y} = \mathbf{b}$. As initial guess we take $\mathbf{y}_0 = \mathbf{0}$. The approximation $\mathbf{y}_k \in \mathcal{K}_k$ is optimal in the sense

that it satisfies

$$\|\mathbf{b} - B\mathbf{y}_k\|_2 = \min_{\mathbf{y}\in\mathcal{K}_k} \|\mathbf{b} - B\mathbf{y}\|_2 \,.$$

The subspace $\mathcal{K}_k = \mathcal{K}_k(B,\mathbf{b}) := \text{span}\{\mathbf{b}, B\mathbf{b}, \ldots, (B)^{k-1}\mathbf{b}\}$ is called the Krylov-space.

Before we discuss the convergence of MINRES we state a result on min-max polynomials based on Chebyshev approximation. For more detail on approximating min-max polynomials using Chebyshev polynomials see Appendix A. To state this result we define the set of normalised polynomials of degree $\leq k$,

$$\Pi_k^1 \;:=\; \{f | f \text{ polynomial, } f(0) = 1, \text{ and degree}(f) \leq k\}. \tag{3.9}$$

Lemma 3.4 *Given $D \subset \mathbb{R}$ compact and $0 \notin D$ then*

$$\min_{f\in\Pi_k^1} \max_{\xi\in D} |f(\xi)| \;\leq\; p\,q^k, \tag{3.10}$$

where $p = 1$ and $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ if $D \subset \mathbb{R}^+$ or $D \subset \mathbb{R}^-$, and $1/p = q = \sqrt{\frac{\kappa-1}{\kappa+1}}$ if $\exists\ \xi_1,\xi_2 \in D$ such that $\xi_1 < 0 < \xi_2$. Here $\kappa := \max_{\xi\in D} |\xi| \,/\, \min_{\xi\in D} |\xi|$.

Proof: See Appendix A, Corollary A.3 part 1 for $D \subset \mathbb{R}+$ or $D \subset \mathbb{R}^-$ and Corollary A.3 part 2 for the other case. □

For the case where $\exists\ \xi_1, \xi_2$ with $\xi_1 < 0 < \xi_2$ the bound is usually written as

$$\min_{f\in\Pi_k^1} \max_{\xi\in D} |f(\xi)| \;\leq\; \left(\frac{\kappa-1}{\kappa+1}\right)^{\left[\frac{k}{2}\right]},$$

however the form in Lemma 3.4 is more convenient later.

When D is the set of eigenvalues of a matrix, say B, then κ is referred to as the condition number of B. If D contains a subset of eigenvalues of B then κ is referred to as the reduced condition number. Obviously if $\kappa \to \infty$ then $q \to 1$ and the rate of convergence deteriorates.

Now consider $\xi_1 \in D$ close to zero, and $D_1 := D\backslash\{\xi_1\}$ being well separated from zero, then $\kappa(D_1) \ll \kappa(D)$. Therefore it might be appropriate to treat ξ_1 separately. If $f \in \Pi_k^1$ then $g(\xi) = f(\xi)(\xi - \xi_1)/\xi_1 \in \Pi_{k+1}^1$ and

$$\max_{\xi\in D} |g(\xi)| \;\leq\; \max_{\xi\in D} \frac{|\xi - \xi_1|}{|\xi_1|} \max_{\xi\in D} |f(\xi)| \,.$$

If most of the eigenvalues have the same sign and only a few are on the other side of the origin, a similar treatment of the few would give $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ instead of $q = \sqrt{\frac{\kappa-1}{\kappa+1}}$. This common technique can, for example, be found in Hackbusch (1994, Section 7.3.6).

Given a set $\Gamma \subset \mathbb{N}_n$ we define

$$Q_\Gamma := \operatorname{diag}(\delta_1(\Gamma), \ldots, \delta_n(\Gamma)) \tag{3.11}$$

where $\delta_j(\Gamma) = 0$ if $j \in \Gamma$ and $\delta_j(\Gamma) = 1$ otherwise.

The convergence result for MINRES which we now present is based on a polynomial bound on the MINRES residual after k (inner) iterations

$$\mathbf{d}_k := \mathbf{b} - B\mathbf{y}_k. \tag{3.12}$$

Lemma 3.5 *Consider MINRES being applied to $B\mathbf{y} = \mathbf{b}$. Let B be non-singular and have the eigenvalue decomposition $B = W\Lambda_B W^T$ with $\Lambda_B = diag(\mu_1, \ldots, \mu_n)$. Then MINRES converges and for all $\Gamma \subset \mathbb{N}_n$ with $1 \in \Gamma$ and $\mu_j \in D$ for $j \notin \Gamma$, there exists $p > 0$ and $q_\Gamma := q \in (0,1)$ as defined in Lemma 3.4, such that*

$$\|\mathbf{d}_k\|_2 \leq (q_\Gamma)^{k-|\Gamma|} p_\Gamma \, |\mu_1|^{-1} \|Q_\Gamma W^T \mathbf{b}\|_2 \tag{3.13}$$

where

$$p_\Gamma := p \max_{\xi \in D} \left(|\mu_1 - \xi| \prod_{j \in \Gamma \setminus \{1\}} \frac{|\mu_j - \xi|}{|\mu_j|} \right). \tag{3.14}$$

Proof: We use that for any $\mathbf{y} \in \mathcal{K}_k$ there exists a polynomial h of degree $\leq k-1$ such that $\mathbf{y}$ can be written as $\mathbf{y} = h(B)\mathbf{b}$ and therefore $\mathbf{d} = \mathbf{b} - Bh(B)\mathbf{b} \in \Pi_k^1$. Hence we obtain

$$\begin{aligned} \|\mathbf{d}_k\|_2 &= \min_{\mathbf{y} \in \mathcal{K}_k} \|\mathbf{b} - B\mathbf{y}\|_2 \\ &= \min_{f \in \Pi_k^1} \|f(B)\mathbf{b}\|_2 \\ &= \min_{f \in \Pi_k^1} \|f(W\Lambda_B W^T)\mathbf{b}\|_2 \\ &= \min_{f \in \Pi_k^1} \|f(\Lambda_B) W^T \mathbf{b}\|_2 \, . \end{aligned}$$

Now set $g(\xi) := \prod_{j \in \Gamma} \frac{\mu_j - \xi}{\mu_j}$ then

$$\begin{aligned} \|\mathbf{d}_k\|_2 &= \min_{f \in P_{k-|\Gamma|}^1} \|f(\Lambda_B) g(\Lambda_B) W^T \mathbf{b}\|_2 \\ &= \min_{f \in P_{k-|\Gamma|}^1} \|f(\Lambda_B) g(\Lambda_B) Q_\Gamma W^T \mathbf{b}\|_2 \\ &\leq \min_{f \in P_{k-|\Gamma|}^1} \|f(\Lambda_B) g(\Lambda_B) Q_\Gamma\| \, \|Q_\Gamma W^T \mathbf{b}\|_2 \end{aligned}$$

$$
\begin{aligned}
&\leq \min_{f \in P^1_{k-|\Gamma|}} \max_{j \in \mathbb{N}_n \setminus \Gamma} \|f(\mu_j) g(\mu_j)\| \, \|Q_\Gamma W^T \mathbf{b}\|_2 \\
&\leq \left(\min_{f \in P^1_{k-|\Gamma|}} \max_{j \in \mathbb{N}_n \setminus \Gamma} \|f(\mu_j)\| \right) \left(\max_{j \in \mathbb{N}_n \setminus \Gamma} \|g(\mu_j)\| \right) \, \|Q_\Gamma W^T \mathbf{b}\|_2 \\
&\leq (q_\Gamma)^{k-|\Gamma|} p_\Gamma \, |\mu_1|^{-1} \|Q_\Gamma W^T \mathbf{b}\|_2 \, .
\end{aligned}
$$

As $q_\Gamma \in (0,1)$ and p_Γ, $|\mu_1|^{-1}$ and $\|Q_\Gamma W^T \mathbf{b}\|_2$ are bounded, therefore $\|\mathbf{d}_k\| \to 0$ as $k \to \infty$, hence MINRES converges. □

The basic idea of this analysis can be found in Hackbusch (1994, Section 7.3.6). van der Vorst and Vuik (1993) use a similar analysis to study the superlinear convergence of GMRES. A more detailed discussion of the convergence of MINRES can be found in Ipsen (1998a). For a derivation of the algorithm based on polynomials see Fischer (1996). As MINRES can be viewed as a special implementation of GMRES, one can apply the convergence results for GMRES, see for example Kelley (1995) and Greenbaum (1997).

A traditional observation for Krylov methods is that the algorithm should not take more than n iterations to find an exact solution. This result can be obtained from Lemma 3.5 by setting $\Gamma = \mathbb{N}_n$, then $\|Q_\Gamma V^T \mathbf{x}^i\|_2 = 0$ and hence $\|\mathbf{res}^i_n\|_2 = 0$.

3.2.2 MINRES as linear solver for shifted systems

In later sections we apply MINRES to sequences of shifted linear systems. Further we will consider the cases where either unpreconditioned MINRES or preconditioned MINRES is applied to such a sequence. To simplify later analysis we discuss both cases here and present a Corollary to Lemma 3.5 applicable to both situations.

Given the linear systems $(A - \sigma^i I)\mathbf{y}^i = \widetilde{\mathbf{b}}^i$, for our application of MINRES, that is as a linear solver in inexact inverse iteration, σ^i will vary only in certain intervals, say $0 < |\lambda_1 - \sigma^i| \leq \frac{1}{2} |\lambda_2 - \lambda_1|$, where $|\lambda_2 - \lambda_1| \leq |\lambda_j - \lambda_1|$ for $j = 3, \ldots, n$. Therefore $\exists$ $a, b > 0$ such that $a \leq |\lambda_j - \sigma^i| \leq b$ for all $j \geq 2$ and $|\lambda_1 - \sigma^i| \leq b$. Then all eigenvalues are bounded and the only eigenvalue which is not separated from the origin is $\lambda_1 - \sigma^i$.

In case of preconditioned MINRES solves we apply to $(A - \sigma^i I)\mathbf{y} = \widetilde{\mathbf{b}^i}$ a symmetric positive preconditioner, say P. As P is spd there exists P_1 such that $P_1 P_1^T = P$, for example a Cholesky preconditioner or the spd square root $P_1 = P^{\frac{1}{2}}$, with $P^{\frac{1}{2}}$ spd and $P^{\frac{1}{2}} P^{\frac{1}{2}} = P$. Such a factorisation is only needed for the theory, the actual algorithm just needs the action of P^{-1} on a vector. However in both cases, unpreconditioned and preconditioned, the system being solved can be written as

$$
P_1^{-1}(A - \sigma^i I) P_1^{-T} \mathbf{z}^i \;=\; P_1^{-1} \widetilde{\mathbf{b}}^i, \tag{3.15}
$$

and $\mathbf{y}^i = P_1^{-T} \mathbf{z}^i$, where $P = I$ for unpreconditioned MINRES. We define the residual

for the ith linear system by

$$\mathbf{d}_k^i := P_1^{-1}\widetilde{\mathbf{b}}^i - P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{z}^i, \tag{3.16}$$

and set $\mathbf{b}^i := P_1^{-1}\widetilde{\mathbf{b}}^i$ while $B^i := P_1^{-1}(A-\sigma^i I)P_1^{-1}$. The following Lemma shows that also in the preconditioned case all eigenvalues but one are bounded and nicely separated from the origin, while the remaining critical eigenvalue is linear in $\lambda - \sigma^i$.

Lemma 3.6 *Let $B \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues $\omega_1(B) \le \ldots \le \omega_n(B)$ and $Z \in \mathbb{R}^{n\times n}$ nonsingular. Further let the eigenvalues of Z^TBZ be ordered such that $\omega_1(Z^TBZ) \le \ldots \le \omega_n(Z^TBZ)$, then*

$$\begin{aligned} & \nu_n^2\omega_j(B) \le \omega_j(Z^TBZ) \le \nu_1^2\omega_j(B), \quad \text{if } \omega_j(B) > 0 \\ \text{and} \quad & \nu_1^2\omega_j(B) \le \omega_j(Z^TBZ) \le \nu_n^2\omega_j(B), \quad \text{if } \omega_j(B) < 0, \end{aligned} \tag{3.17}$$

where ν_1 the largest and ν_n is the smallest singular value of Z.

Proof: See proof of Sylvester's Inertia Theorem, as given in Golub and van Loan (1996, Theorem 8.1.17). □

Given a set $\Gamma \subset \mathbb{N}_n$ with $1 \in \Gamma$ then denote by

$$\begin{aligned} a &:= \nu_n^2 \min_{j\in\mathbb{N}\setminus\Gamma}(|\lambda_j - \lambda_1| - \frac{1}{2}|\lambda_2 - \lambda_1|) \quad \text{and} \\ b &:= \nu_1^2 \max_{j\in\mathbb{N}\setminus\Gamma}(|\lambda_j - \lambda_1| + \frac{1}{2}|\lambda_2 - \lambda_1|), \end{aligned}$$

where ν_1 is the largest and ν_n the smallest singular value of P_1^{-1}. In the case where $\lambda_j > \lambda_1$ for all $j \notin \Gamma$ then define $D_\Gamma := [a,b]$ and in case $\lambda_j < \lambda_1$ for all $j \notin \Gamma$ then define $D_\Gamma := [-b,-a]$ while $D_\Gamma := [-b,-a] \cup [a,b]$ otherwise. Then Lemma 3.6 states that $\mu_j^i \in D_\Gamma$ for $j \notin \Gamma$ and D_Γ independent of the shift, so that the constants q_Γ and p_Γ as given by Lemma 3.5 are independent of σ^i. The constants q_Γ and p_Γ might be improved using $D_\Gamma^i := \{\mu_j^i | j \notin \Gamma\}$, however it is convenient to have q_Γ and p_Γ independent of σ^i.

Corollary 3.7 *Consider MINRES being applied to the linear system $P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{z}^i = \mathbf{b}^i$. Denote the eigenvalues of A by $\lambda_1,\ldots,\lambda_n$ and assume $0 < |\lambda_1 - \sigma^i| \le \frac{1}{2}|\lambda_2 - \lambda_1|$ and let P_1 be nonsingular, then MINRES converges and for all $\Gamma \subset \mathbb{N}_n$ there exist $p_\Gamma > 0$ and $q_\Gamma \in (0,1)$ such that for the residual $\mathbf{d}_k^i := \mathbf{b}^i - P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{z}_k^i$ the bound*

$$\|\mathbf{d}_k^i\|_2 \le (q_\Gamma)^{k-|\Gamma|} p_\Gamma |\lambda_1 - \sigma^i|^{-1} \chi^i, \tag{3.18}$$

holds for all i. Here $\chi^i := \|Q_\Gamma(W^i)^T\mathbf{b}^i\|_2$ and W^i denotes the matrix of eigenvectors of $P_1^{-1}(A-\sigma^i I)P_1^{-T}$.

Proof: For all i the conditions of Lemma 3.5 are satisfied and the result is obtained with $D = D_\Gamma$ independent of i. □

The bound is rather pessimistic which is due to the definition of D_Γ. Often one has more knowledge about the preconditioner and can provide therefore better bounds for the preconditioned eigenvalues. Such additional knowledge does not change the character of the bound but improves the values for p_Γ and q_Γ. Another approach is to set

$$D_\Gamma := \left\{ \mu_j | j \notin \Gamma, \text{ and } \exists \sigma \text{ with } 0 < |\lambda_1 - \sigma| \leq \frac{1}{2} |\lambda_2 - \lambda_1|, \text{ such that } B = P_1^{-1}(A - \sigma I)P_1^{-T} \text{ has the eigenvalue } \mu_j \right\}, \tag{3.19}$$

which is compact and separated from the origin. Finally one can restrict this D_Γ to those σ^i which appear in the practical algorithm. This final step makes the comparison between different variations of inexact inverse iteration cumbersome, however it can be used to verify the descriptive quality of Corollary 3.7.

3.3 Efficiency for unpreconditioned MINRES solves

In this section we want to determine which choice of parameter makes inexact inverse iteration using unpreconditioned MINRES (Invit+MINRES) efficient. Previously, in Section 2.2 we assumed for the convergence analysis that the linear solver is capable of providing a solution satisfying the residual condition $\|\mathbf{res}\| \leq \tau$. Now by considering MINRES as linear solver we have to show that MINRES is capable of providing a solution satisfying the residual constraint in order to prove convergence for Invit+MINRES. Essentially this has been done in the previous section, but we didn't state this explicitly. The main aim of this section is to determine which choice of parameters makes Invit+MINRES efficient we will bound the cost of Invit+MINRES. Therefore we define the minimal number of inner iterations used to satisfy the residual condition $\|\mathbf{res}\| \leq \tau$ as a measure for the cost of a linear solve. Similarly we define the overall cost of Invit+MINRES as the total number of inner iterations. For both cost measures we derive a posteriori bounds which link the cost of a linear solve or an eigenvalue calculation to the progress achieved. Based on the bounds we show how to choose the parameters to obtain an efficient method. Also we make some remarks on the practical use of the obtained results.

3.3.1 Measures for costs

As the most expensive operation in unpreconditioned MINRES is a matrix vector product we can use the number of matrix vector products (equivalently the number of inner iterations) to measure the cost of a linear solve using MINRES. Other costs are

storage, orthonormalisation, and updating the solution. Since, the amount of storage is typically fixed and the remaining costs are linear in the number of inner iterations, the number of inner iterations is an appropriate measure for the cost of a MINRES solve.

MINRES can be stopped using different stopping conditions, for example after a prescribed number of inner iterations. Later in Section 3.6.1 we discuss the stopping condition recently suggested by Simoncini and Eldén (2002). In the convergence theory in Chapter 2 we used the condition $\|\mathbf{res}_k^i\| \leq \tau^i$ to measure the quality of the iterative linear solve, hence we regard this as the appropriate stopping condition for MINRES. Hence we use the following Definition for the number of inner iterations, which we later regard as the cost of a linear solve.

Definition 3.8 *For the system $B\mathbf{y} = \mathbf{b}$ let $\tau > 0$ be the relative accuracy requirement, $\|\mathbf{res}(\widetilde{\mathbf{y}})\|_2 \leq \tau$ for the approximate solution $\widetilde{\mathbf{y}}$, that is, $\widetilde{\mathbf{y}}$ is acceptable if the residual $\mathbf{res}(\widetilde{\mathbf{y}}) := \mathbf{b} - B\widetilde{\mathbf{y}}$ satisfies $\|\mathbf{res}(\widetilde{\mathbf{y}})\|_2 \leq \tau \|\mathbf{b}\|$. Then define $\mathcal{L} \in \mathbb{N}_0$ as the minimal number of inner iterations needed by MINRES such that the accuracy required is achieved, that is*

$$\|\mathbf{res}_{\mathcal{L}}\|_2 \leq \tau \|\mathbf{b}\|_2 \quad \textit{and} \quad \|\mathbf{res}_k\|_2 > \tau \|\mathbf{b}\|_2 \ \ \forall 0 \leq k < \mathcal{L}.$$

Based on this definition we define the overall costs. As the most costly part of inexact inverse iteration by far is the linear solve we neglect the costs which arise from the remaining steps in the outer method. Hence we use as a measure for the total costs, the total number of inner iterations.

Definition 3.9 *Suppose we are given a matrix A, a starting vector $\mathbf{x}^0$, a sequence of shifts (σ^i), and a sequence of accuracy requirements $(\tau^i) \in \mathbb{R}^+$ for the linear solves. Further assume that for all iteration $i \geq 0$ $\mathcal{L}^i$ exists, where $\mathcal{L}^i$ as in Definition 3.8. Then define the total cost $\mathcal{T}$ as the sum of all inner iterations needed to improve t^0 by a factor $10^{-\gamma}$, that is $\mathcal{T} := \sum_{i=0}^{\mathcal{N}-1} \mathcal{L}^i$, where $\mathcal{N}$ is the number of outer iterations defined in Definition 3.1.*

While the definition for the number of outer iterations, Definition 3.1, is independent of the linear solver, Definitions 3.8 and 3.9 depend on the linear solver.

3.3.2 Efficiency analysis

As mentioned in the beginning of this section we have not proven the convergence of Invit+MINRES. Combining Lemma 2.1 and Definition 3.8 it remains to prove that for each outer iteration there is a $\mathcal{L}^i$. The following theorem will provide this convergence result and also states a posteriori bounds on $\mathcal{L}^i$ and $\mathcal{T}$.

Theorem 3.10 *Consider inexact inverse iteration, defined by Algorithm 2, using unpreconditioned MINRES applied to $A \in \mathbb{R}^{n\times n}$, symmetric. Assume the conditions of Theorem 2.1 are satisfied. Then convergence is obtained (i.e. $t^i \to 0$, $\varrho(\mathbf{x}^i) \to \lambda_1$ and $\mathbf{x}^i \to \mathbf{v}_1$). Further $\mathcal{L}^i$ as defined in Definition 3.8 exists for all i, and if for some $C_5 \in \mathbb{R}^+$ and all i, $|s^i| \leq C_5 \, \|\mathbf{res}^i_{\mathcal{L}^i}\|_2$ then there exists $p_\Gamma \in \mathbb{R}^+$ and $q_\Gamma \in (0,1)$ such that*

$$\mathcal{L}^i \;<\; 1 + |\Gamma| + \frac{C_6}{\log((q_\Gamma)^{-1})} + \frac{\log \frac{t^i}{t^{i+1}}}{\log((q_\Gamma)^{-1})}, \tag{3.20}$$

where

$$C_6 \;:=\; \log \frac{4(1+C_5)p_\Gamma}{|\lambda_2 - \lambda_1|\,(1-C_3)}. \tag{3.21}$$

If t^0 should be improved by a factor $10^{-\gamma}$, with $\gamma > 0$ then

$$\mathcal{T} \;\leq\; \frac{\gamma \log 10 + \log \frac{t^0 10^{-\gamma}}{t^{\mathcal{N}}}}{\log((q_\Gamma)^{-1})} + \mathcal{N}\left(1 + |\Gamma| + \frac{C_6}{\log((q_\Gamma)^{-1})}\right). \tag{3.22}$$

Proof: We start by proving the existence of $\mathcal{L}^i$, as defined in Definition 3.8. Then we use the bound from Corollary 3.7 to prove the bound on $\mathcal{L}^i$ and finally apply Definition 3.9 to obtain the bound on the total cost $\mathcal{T}$.

For each outer iteration i the conditions of Corollary 3.7 are satisfied and therefore MINRES converges. Hence $\exists\ \mathcal{L}^i \in \mathbb{N}$ for each outer iteration and as the conditions of Theorem 2.1 are satisfied, Algorithm 2 converges towards the desired solution.

We now prove the bound on $\mathcal{L}^i$. We use the MINRES residual bound from Lemma 3.5 together with the definition of $\mathcal{L}^i$ to obtain

$$\tau^i \;<\; \|\mathbf{res}^i_{\mathcal{L}^i-1}\|_2 \;\leq\; {q_\Gamma}^{\mathcal{L}^i - 1 - |\Gamma|} p_\Gamma \, |\lambda_1 - \sigma^i|^{-1} \, \chi^i. \tag{3.23}$$

The value χ^i was defined in Corollary 3.7 as

$$\chi^i \;=\; \|Q_\Gamma V^T \mathbf{x}^i\| \;\leq\; \|Q_1 V^T \mathbf{x}^i\| \;=\; \|\mathbf{v}_1^T \mathbf{x}^i\| \;=\; |s^i|,$$

where Q_1 is Q_Γ for $\Gamma = \{1\}$. By rearranging and taking logarithms we gain from (3.23)

$$\mathcal{L}^i \;<\; 1 + |\Gamma| + \log\left(\frac{p_\Gamma t^i}{\tau^i \, |\lambda_1 - \sigma^i|}\right) / \log((q_\Gamma)^{-1}). \tag{3.24}$$

To link this with the outer convergence we use the one-step bound (2.18)

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1 - \sigma^i|\,|s^i| + \|\mathbf{res}^i\|_2}{|\lambda_2 - \sigma^i|\,|c^i| - \|\mathbf{res}^i\|_2} \\ &\leq 4\,|\lambda_2 - \lambda_1|^{-1}\,(1 - C_3)^{-1}(1 + C_5)\,|\lambda_1 - \sigma^i|\,\tau^i, \end{aligned} \tag{3.25}$$

hence by further rearranging

$$|\lambda_1 - \sigma^i|^{-1}\,(\tau^i)^{-1} \leq 4\,|\lambda_2 - \lambda_1|^{-1}\,(1 - C_3)^{-1}(1 + C_5)(t^{i+1})^{-1}. \tag{3.26}$$

Combining (3.24) with (3.26) we obtain the bound on $\mathcal{L}^i$

$$\mathcal{L}^i < 1 + |\Gamma| + \log\left(\frac{4(1 + C_5)p_\Gamma t^i}{|\lambda_2 - \lambda_1|\,(1 - C_3)t^{i+1}}\right) / \log((q_\Gamma)^{-1}).$$

Finally we use the definition of $\mathcal{T}$, Definition 3.9, and gain

$$\begin{aligned} \mathcal{T} &= \sum_{i=0}^{\mathcal{N}-1} \mathcal{L}^i \\ &\leq \sum_{i=0}^{\mathcal{N}-1} \frac{\log \frac{t^i}{t^{i+1}}}{\log((q_\Gamma)^{-1})} + \sum_{i=0}^{\mathcal{N}-1} 1 + |\Gamma| + \frac{\log \frac{4(1 + C_5)p_\Gamma}{|\lambda_2 - \lambda_1|\,(1 - C_3)}}{\log((q_\Gamma)^{-1})} \\ &\leq \frac{\log \frac{t^0}{t^{\mathcal{N}}}}{\log((q_\Gamma)^{-1})} + \mathcal{N}\left(1 + |\Gamma| + \frac{C_6}{\log((q_\Gamma)^{-1})}\right), \end{aligned}$$

from which we obtain (3.22). □

The bounds (3.20) and (3.22) in Theorem 3.10 are a-posteriori bounds, as t^{i+1} is used on the right-hand side which is only available after the linear solve has been carried out using $\mathcal{L}^i$ iterations. Despite the fact that the bound is a-posteriori its nature is more like that of an a-priori bound. The only two a-priori unknown variables on the right-hand side are $t^{\mathcal{N}}$ and $\mathcal{N}$. While the first can be controlled by including an additional stopping condition in the linear solver, the second is somehow (we discuss this later in more detail) a-priori controlled by the choice of the method. An a priori bound for $\mathcal{L}^i$ is inequality (3.24). We are aware that the common technique of presenting bounds on the number of iterations is different. Often such a bound is given in the sense that if $k \geq \ldots$ then convergence is achieved. This would extend to, if $\mathcal{T} \geq \ldots$ and all $\mathcal{L}^i \geq \ldots$ then the convergence for inexact inverse iteration is achieved. Results of this form can be found in Berns-Müller et al. (2003).

As we will see in Section 3.8 where we illustrate the results considering a few numerical examples, the bound on the number of inner iterations $\mathcal{L}^i$ (3.20) is not sharp but it mirrors the underlying behaviour sufficiently well.

In the bound for the total cost $\mathcal{T}$, (3.22), the numerator of the first term is related to the progress the algorithm has achieved on t^0 after $\mathcal{N}$ iterations. The first of the two terms in this numerator corresponds to the task, while the second corresponds to what has been achieved additionally to the asked convergence level. If one wants to reduce this additional achievement, then an additional stopping condition for the inner iteration is appropriate. We discuss this later in Section 3.7 in more detail.

3.3.3 Mesh dependency of the costs, a theoretical example

Based on a simple example we will explain how the number of inner iterations depends on the mesh-size h for a PDE problem.

If we think of A as a discretisation of a second order differential operator, then some eigenvalues of A will depend on the mesh-size h. With a change in the mesh-size the parameters p_Γ and q_Γ will alter and therefore $\mathcal{L}^i$ will vary with h. To understand the dependence of p_Γ, q_Γ, $\mathcal{L}^i$ and $\mathcal{T}$ on h we look at one specific example.

Consider the Poisson eigenvalue problem

$$-(u_{xx} + u_{yy}) = \lambda u \tag{3.27}$$

on the unit square with boundary data $u(0,y) = u(1,y) = u(x,0) = u(x,1) = 0$. Let A be derived by discretising (3.27) using a second order, central finite difference scheme on a uniform square mesh. By doing so we derive a standard symmetric eigenvalue problem $A\mathbf{x} = \lambda\mathbf{x}$. The eigenvalues of A are given by

$$\frac{1}{h^2}\left(4 - 2\cos(\frac{j}{m}\pi) - 2\cos(\frac{l}{m}\pi)\right) \text{ for } 1 \le j,l \le m-1,$$

where $h = 1/m$. For more detail see for example Strang (1986, p.456 and p.571). To simplify the calculation we assume that $h \ll \pi$ and use that $\lambda_1 \to 2\pi^2$ and $\lambda_2 \to 5\pi^2$, for $h \to 0$ while $|\lambda_n - \lambda_1| \approx 8h^{-2}$. Further we choose $\Gamma = \{1\}$. Then we can approximate the reduced condition number by

$$\kappa_\Gamma \approx \frac{|\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1|} \approx \frac{8}{3\pi^2 h^2}$$

and gain for q_Γ according to Lemma (3.5) $q_\Gamma \approx (2\sqrt{2} - \pi h\sqrt{3})/(2\sqrt{2} + \pi h\sqrt{3})$. Hence for h small enough we gain $\log((q_\Gamma)^{-1}) \approx 2\pi h\sqrt{3}/(2\sqrt{2} - \pi h\sqrt{3})$. Now using the definition of p_Γ, given in (3.14), we gain $p_\Gamma = |\lambda_n - \lambda_1| \approx 8h^{-2}$. Applying p_Γ and q_Γ to the bound for $\mathcal{L}^i$, (3.20), gives

$$\mathcal{L}^i < 2 + \frac{\log\frac{4(1+C_5)}{1-C_3} + \log\frac{p_\Gamma}{|\lambda_2 - \lambda_1|} + \log\frac{t^i}{t^{i+1}}}{\log((q_\Gamma)^{-1})}$$

$$\begin{aligned} &\approx 2 + \frac{2\sqrt{2} - 3\pi h\sqrt{3}}{2\pi h\sqrt{3}} \left(\log \frac{4(1+C_5)}{1-C_3} + \log \frac{t^i}{t^{i+1}} - \log 3\pi^2 - \log h^{-2} \right) \\ &= O\left(\frac{1}{h} \log \frac{1}{h} \right). \end{aligned}$$

This dependency of $\mathcal{L}^i$ on h is a typical result for Conjugate Gradient type convergence as we expect for MINRES applied to an extreme eigenvalue. The order of $\mathcal{L}^i$ changes if λ_1 is an interior eigenvalue and there is no small set Γ such that all eigenvalues corresponding to $\mathbb{N}\backslash\Gamma$ are to one side of λ_1. In such a case the dependency gets worse, $\mathcal{L}^i = O(h^{-2}\log((h)^{-1}))$.

We now look at the total work $\mathcal{T}$, for which we consider the case where the constants C_1, C_2, and C_3 in Theorem 2.1 are independent of the meshparameter h. Then $\mathcal{N}^*$, as defined in (3.1), is independent of h and therefore $\mathcal{T} = O(h^{-1}\log(h^{-1}))$.

However we do not produce any tests to confirm these theoretical results.

3.3.4 Optimal strategy

As we do not know $\mathcal{T}$ itself we can base our judgement only on the bound of $\mathcal{T}$. However such an approach can be understood as limiting the worst case performance. In practice this worst case approach is sensible as it points the way to reliable methods.

In the convergence theory in Chapter 2 we have seen that $\|\mathbf{res}^i\| \,/\, |s^i| \to 0$ does not give any benefit over $\|\mathbf{res}^i\| \,/\, |s^i| = const.$ Further in the efficiency result, Theorem 3.10, a bound $(C_5) \geq |s^i| \,/\, \|\mathbf{res}^i\|$ is required. As $\mathcal{T} \propto \log(1 + C_5)$ large values for C_5 might be avoided in order to reduce $\mathcal{T}$.

Lemma 3.11 *Consider the conditions of Theorem 3.10 being satisfied. For γ sufficiently large it is optimal to choose σ^i and τ^i such to minimize $\mathcal{N}$ while keeping $\|\mathbf{res}^i\| \,/\, |s^i|$ large.*

Proof: For γ large enough the discrete nature of $\mathcal{T}$ and $\mathcal{N}$ can be neglected. Then from Theorem 3.10 we gain that $\mathcal{T}$ is linear in $\mathcal{N}$. Next from Lemma 3.2 we gain $\mathcal{N} \leq 1 + [\mathcal{N}^*]$ and from Lemma 3.3 that $\mathcal{N}^*$ is decreasing in $\alpha + \beta$. Thus the bound on $\mathcal{T}$ is minimised for $\alpha + \beta$ maximal, but as $\beta \in [0, 1]$ this reduces to α maximal and $\beta = 1$. As the influence of C_5 is more dominant than the one of C_2 it is better to keep $\|\mathbf{res}^i\| \,/\, \|s^i\|$ large. □

Corollary 3.12 *Consider the conditions of Theorem 3.10 being satisfied. Assume the Rayleigh quotient is the best shift then for γ sufficiently large the RQI with decreasing tolerance ($\|\mathbf{res}^i\| = O(|s^i|)$), is the most efficient method.*

We remark that this is a theoretical result assuming ∞-precision arithmetic. In practice, due to γ not large enough, the cubically convergence RQI with decreasing

tolerance might need the same number of outer iterations as the quadratically converging RQI with fixed shift. In such a case we would expect the quadratically converging method to be at least competitive. Based on these practical effects, the theoretical advantage of cubically converging methods might not be observed in practice, see Section 2.3.3. We demonstrated such effects earlier in Section 2.3 and give more evidence in Section 3.8. However to ensure a small number of total inner iteration $\mathcal{T}$ one has to reduce the number of outer iterations. This can be achieved by shifting towards the singularity using the Rayleigh quotient.

3.4 Efficiency for preconditioned MINRES solves

The methods covered in this section represent the standard approach of using preconditioned MINRES in inexact inverse iteration. As we observe later, these methods have, at least in theory, larger convergence areas than the approach from Simoncini and Eldén (2002), which we discuss later in Section 3.6. However the methods discussed in this Section are not as efficient as the one from Simoncini and Eldén (2002). In order to understand the difference and to appreciate the advantage the later discussed methods have we give a brief discussion here for the standard approach of using inexact inverse iteration with preconditioned MINRES. The techniques we apply are the same as in Section 3.3, however, the results for $\mathcal{L}^i$ and $\mathcal{T}$ differ in their structure. As this different structure is inferior, the results in this chapter motivate the discussion on the inverse correction method and the approach of Simoncini and Eldén (2002) in the following two sections.

Theorem 3.13 *Consider inexact inverse iteration, defined by Algorithm 2, using preconditioned MINRES with a positive definite preconditioner P applied to $A \in \mathbb{R}^{n\times n}$, symmetric. Assume the conditions of Theorem 2.1 are satisfied so that convergence is obtained. Further, for all i, $\mathcal{L}^i$, as defined in Definition 3.8, exists and if for some $C_5 \in \mathbb{R}^+$ and all i, $|s^i| \leq C_5 \|\mathbf{res}^i_{\mathcal{L}^i}\|_2$ then there exists $p_\Gamma \in \mathbb{R}^+$ and $q_\Gamma \in (0,1)$ as defined in Lemma 3.5 such that*

$$\mathcal{L}^i \;<\; 1+|\Gamma| + \frac{C_6}{\log((q_\Gamma)^{-1})} + \frac{\log\frac{1}{t^{i+1}}}{\log((q_\Gamma)^{-1})}, \tag{3.28}$$

where

$$C_6 \;:=\; \log\frac{4(1+C_5)p_\Gamma\,\nu_1}{|\lambda_2-\lambda_1|\,(1-C_3)\nu_n} / \log((q_\Gamma)^{-1}), \tag{3.29}$$

while ν_1 is the largest and ν_n the smallest singular value of P_1. If t^0 should be improved

by a factor $10^{-\gamma}$, *with* $\gamma > 0$ *then*

$$\mathcal{T} \leq \frac{\sum_{i=0}^{\mathcal{N}-1} \log \frac{1}{t^{i+1}}}{\log((q_\Gamma)^{-1})} + \mathcal{N}\left(1 + |\Gamma| + \frac{C_6}{\log((q_\Gamma)^{-1})}\right). \tag{3.30}$$

The proof is similar in technique to the proof of Theorem 3.10.

Proof: We start the proof by using Theorem 2.1 and Corollary 3.6 to prove convergence and the existence of $\mathcal{L}^i$, as defined in Definition 3.8. Then we use the bound from Lemma 3.6 to prove the bound on $\mathcal{L}^i$ and finally apply Definition 3.9 to obtain the bound on the total cost $\mathcal{T}$.

For each outer iteration i the conditions of Corollary 3.7 are satisfied and therefore preconditioned MINRES converges. Hence for all i there exists $\mathcal{L}^i \in \mathbb{N}$ as in Definition 3.8, and as the conditions of Theorem 2.1 are satisfied Algorithm 2 converges towards the desired eigenpair.

We now prove the bound on $\mathcal{L}^i$. To do so we observe that

$$\begin{aligned} \mathbf{res}_k^i &= \mathbf{x}^i - (A - \sigma^i I)\mathbf{y}_k^i \\ &= P_1\left(P_1^{-1}\mathbf{x}^i - P_1^{-1}(A - \sigma^i I)P_1^{-T}\mathbf{z}_k^i\right) \\ &= P_1\mathbf{d}_k^i. \end{aligned}$$

Now we use the MINRES residual bound from Corollary 3.6 together with the definition of $\mathcal{L}^i$ to obtain

$$\begin{aligned} \tau^i < \|\mathbf{res}_{\mathcal{L}^i-1}^i\|_2 &\leq \|P_1\|\,\|\mathbf{d}_k^i\| = \nu_1\,\|\mathbf{d}_k^i\| \\ &\leq \nu_1 {q_\Gamma}^{\mathcal{L}^i-1-|\Gamma|} p_\Gamma\,|\lambda_1 - \sigma^i|^{-1}\,\chi^i. \end{aligned} \tag{3.31}$$

The value of χ^i is defined in Corollary 3.6 as $\chi^i = \|Q_\Gamma(W^i)^T\mathbf{b}^i\|$, where W^i is the matrix of eigenvectors of the preconditioned system $P_1^{-1}(A - \sigma I)P_1^{-T}$, and $\mathbf{b}^i = P_1^{-1}\mathbf{x}^i$. As Q_Γ and W^i orthogonal we can bound $\chi^i \leq \|\mathbf{b}^i\| \leq \|P_1^{-1}\| = (\nu_n)^{-1}$. By rearranging and taking logarithms we gain from (3.31)

$$\mathcal{L}^i < 1 + |\Gamma| + \log\left(\frac{p_\Gamma\,\nu_1}{\tau^i\,|\lambda_1 - \sigma^i|\,\nu_n}\right) / \log((q_\Gamma)^{-1}). \tag{3.32}$$

To link this with the outer convergence we use the one-step bound (2.18)

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1 - \sigma^i|\,|s^i| + \|\mathbf{res}^i\|_2}{|\lambda_2 - \sigma^i|\,|c^i| - \|\mathbf{res}^i\|_2} \\ &\leq 4\,|\lambda_2 - \lambda_1|^{-1}\,(1 - C_3)^{-1}(1 + C_5)\,|\lambda_1 - \sigma^i|\,\tau^i, \end{aligned}$$

hence by rearranging

$$|\lambda_1 - \sigma^i|^{-1} (\tau^i)^{-1} \leq 4 |\lambda_2 - \lambda_1|^{-1} (1 - C_3)^{-1}(1 + C_5)(t^{i+1})^{-1}.$$

Together with (3.32) we obtain

$$\mathcal{L}^i < 1 + |\Gamma| + \log\left(\frac{4(1 + C_5)p_\Gamma \nu_1}{|\lambda_2 - \lambda_1| (1 - C_3)\nu_n t^{i+1}}\right) / \log((q_\Gamma)^{-1}).$$

Now we use the definition of $\mathcal{T}$, Definition 3.9, and gain

$$\begin{aligned}
\mathcal{T} &= \sum_{i=0}^{\mathcal{N}-1} \mathcal{L}^i \\
&\leq \sum_{i=0}^{\mathcal{N}-1} \frac{\log \frac{1}{t^{i+1}}}{\log((q_\Gamma)^{-1})} + \sum_{i=0}^{\mathcal{N}-1} 1 + |\Gamma| + \frac{\log \frac{4(1 + C_5)p_\Gamma \nu_1}{|\lambda_2 - \lambda_1| (1 - C_3)\nu_n}}{\log((q_\Gamma)^{-1})} \\
&\leq \sum_{i=0}^{\mathcal{N}-1} \frac{\log \frac{1}{t^{i+1}}}{\log((q_\Gamma)^{-1})} + \mathcal{N}\left(1 + |\Gamma| + \frac{C_6}{\log((q_\Gamma)^{-1})}\right).
\end{aligned}$$

□

One major difference between Theorems 3.10 and 3.13 with respect to $\mathcal{L}^i$ is that for the preconditioned case, Theorem 3.13, the right-hand side is no longer linear in t^i. This difference then carries over to the bound for the total cost $\mathcal{T}$. This effect is not caused by the analysis, but by the fact that preconditioning is carried out as demonstrated by tests. Later in Section 3.8 we show that the gap between the bound and the data is small, see Example 3.2 and Figure 3-2, however the bound is not sharp in the mathematical sense.

Another difference between Theorems 3.10 and 3.13 is that the values for p_Γ and q_Γ change by preconditioning, actually this is the essence of preconditioning. Again it is optimal to use the RQ as shift (in absence of a better approximation of the desired eigenvalue). In contrast to the unpreconditioned case the bound on $\mathcal{L}^i$ agrees not only in essence but can be observed in practice.

3.5 An alternative approach

In the previous section we observed that preconditioning changes the bound on the number of inner iterations. To recapture the structure of the bounds as in the unpreconditioned case while p_Γ and q_Γ benefit from preconditioning we exploit here one alternative approaches for obtaining $\mathbf{y}^i$. Our main focus will be how the convergence theory from Chapter 2 and the efficiency analysis of Sections 3.3 and 3.4 can be extended to this alternative approach. The important difference to inexact inverse iteration as

Algorithm 3: Inverse Correction Method

Given $\mathbf{x}^0$,
For $i = 0, 1, 2, \ldots$

- Choose σ^i and τ^i,
- Calculate $\varrho^i := (\mathbf{x}^i)^T A\mathbf{x}^i$ and $\mathbf{r}^i := (A - \varrho^i I)\mathbf{x}^i$,
- Solve $(A - \sigma^i I)\mathbf{z}^i = \mathbf{r}^i$ such that $\|\mathbf{r}^i - (A - \sigma^i I)\mathbf{z}^i\| \leq \tau^i$,
- Set $\mathbf{y}^i = \mathbf{x}^i - \mathbf{z}^i$,
- Update $\mathbf{x}^{i+1} = \mathbf{y}^i / \|\mathbf{y}^i\|$,
- Test for convergence

discussed so far is that only a correction equation will be solved in order to update the current approximation. In practice methods refining a current approximation by corrections, like the inverse correction method, are often preferred due to their robustness with iterative linear solvers and their efficiency.

The inverse correction method by Rüde and Schmid (1995) and Zaslavsky (1995) is designed to overcome stagnation when inexact inverse iteration is combined with a multigrid solver. The idea is simply to rearrange the solve in inverse iteration, such that only a correction equation has to be solved. In the case of exact linear solves, the two methods are the same. However if we consider iterative solves then this will not necessarily be the case. A similar algorithm has been proposed by Neumaier (1985) to obtain eigenpair approximation of high accuracy from linear and non linear eigenvalue problems. Also, the algorithm of Golub and Ye (2000) can be viewed as the inverse correction method, we explain this towards the end of this section. A more thorough discussion of Golub and Ye (2000) is presented in Chapter 4.

Before we analyse the relation of inverse iteration and inverse correction we discuss why this method might be attractive. First we observe that if $\varrho^i \to \lambda_1$, then $\mathbf{r}^i \to (A - \lambda_1 I)\mathbf{u}^i \perp \mathbf{v}_1$. So the dominant part of the error direction will be orthogonal to the sought eigenvector, this is often thought to be beneficial in solving ill conditioned linear systems, see, for example, Brown and Walker (1997). Specially for the case of multigrid as linear solver it is argued, for example in Rüde and Schmid (1995) that these correction equations are easier to solve as multigrid performs as if this critical eigenvalue does not exist. Like the inverse correction method, the Jacobi-Davidson-

Method (for a review see Sleijpen and van der Vorst (2000)) uses correction equations. In case of the Jacobi-Davidson method the linear system matrix is multiplied from both sides with the projection matrix $I - \mathbf{x}^i(\mathbf{x}^i)^T$ in order to improve the performance of the linear solver. So again the idea is to make the linear solve less costly. However, as we show later in Section 3.8 the inverse correction method suffers either from slow or from erratic convergence when MINRES is used a linear solver.

We start by analysing the convergence of the Inverse Correction Method as given in Algorithm 3. We will do this by linking Algorithm 3 to inexact inverse iteration. Then we present a result on the efficiency of the Inverse Correction Method.

Convergence

In order to distinguish between the variables of both algorithms, we write those from inexact inverse iteration with a tilde and those from Inverse Correction with the subindex $_{ICM}$. In inexact inverse iteration, see Algorithm 1, the next vector is given by

$$\widetilde{\mathbf{x}}^{i+1} = \frac{\widetilde{\mathbf{y}}^i}{\|\widetilde{\mathbf{y}}^i\|_2} = \frac{(A - \widetilde{\sigma}^i I)^{-1}(\widetilde{\mathbf{x}}^i - \widetilde{\mathbf{res}}^i)}{\|\widetilde{\mathbf{y}}^i\|_2}.$$

Now we consider the same iteration for the inverse correction method, as defined by Algorithm 3, then

$$\begin{aligned}
\mathbf{x}_{ICM}^{i+1} &= \frac{\mathbf{x}_{ICM}^i - \mathbf{z}_{ICM}^i}{\|\mathbf{y}_{ICM}^i\|_2} \\
&= \frac{1}{\|\mathbf{y}_{ICM}^i\|_2}\Big(\mathbf{x}_{ICM}^i - (A - \sigma_{ICM}^i I)^{-1}(\mathbf{r}_{ICM}^i + \mathbf{res}_{ICM}^i)\Big) \\
&= \frac{\varrho_{ICM}^i - \sigma_{ICM}^i}{\|\mathbf{y}_{ICM}^i\|_2}(A - \sigma_{ICM}^i I)^{-1}\left(\mathbf{x}_{ICM}^i + \frac{1}{\varrho_{ICM}^i - \sigma_{ICM}^i}\mathbf{res}_{ICM}^i\right) \quad (3.33)
\end{aligned}$$

If $\widetilde{\mathbf{x}}^i = \delta\mathbf{x}_{ICM}^i$, with $|\delta| = 1$ then $\varrho_{ICM}^i = \widetilde{\varrho}^i =: \varrho$. If additionally $\sigma_{ICM}^i = \widetilde{\sigma}^i =: \sigma^i$ and $\widetilde{\mathbf{res}}^i = -\text{sign}((\widetilde{\mathbf{x}}^i)^T\mathbf{x}_{ICM}^i))(\varrho^i - \sigma^i)^{-1}\mathbf{res}_{ICM}^i$, then

$$\mathbf{x}_{ICM}^{i+1} = \text{sign}(\varrho^i - \sigma^i)\widetilde{\mathbf{x}}^{i+1}, \quad (3.34)$$

due to $\|\mathbf{x}_{ICM}^{i+1}\|_2 = \|\widetilde{\mathbf{x}}^{i+1}\|_2$, hence we gain that $\mathbf{x}_{ICM}^{i+1}$ and $\widetilde{\mathbf{x}}^{i+1}$ span the same subspace. Therefore we obtain for the case of exact solves that $\widetilde{\mathbf{x}}^{i+1} = \pm\mathbf{x}_{ICM}^{i+1}$ when $\sigma_{ICM}^i = \widetilde{\sigma}^i$. As the orientation has no effect on the convergence one can say that both algorithms are equivalent when exact solves are used.

Based on above observation we will prove the convergence for the Inverse Correction Method. As both algorithms use inexact solves the next iteration is not a priori defined. In order to overcome this we will look at all solutions x_{ICM}^{i+1} permitted for Inverse Correction for a given $\mathbf{x}^i$, τ_{ICM}^i and σ^i. And we will show that all such solutions are

also permitted for inexact inverse iteration for given $\mathbf{x}^i$, $\widetilde{\tau}^i$ and σ^i.

Lemma 3.14 *Assume that the inverse correction method, defined by Algorithm 3, is applied to $A \in \mathbb{R}^{n\times n}$, A symmetric. Assume $\exists C_1, C_2, \alpha \in \mathbb{R}^+$ and $\beta \in [0,1]$ and $C_3 \in [0,1)$ such that for all $\mathbf{x}^i = c^i\mathbf{v}_1 + s^i\mathbf{u}^i$ the shift satisfies*

$$|\lambda_1 - \sigma^i| \;\leq\; \min\{C_1\,|s^i|^\alpha, \frac{1}{2}\,|\lambda_2 - \lambda_1|\}$$

and $\sigma^i \neq \varrho^i$, and further that the residual satisfies

$$\|\mathbf{res}^i\|_2 \;\leq\; |\varrho^i - \sigma^i|\,\min\{C_2\,|s^i|^\beta, C_3\,|c^i|\} \tag{3.35}$$

for $\alpha + \beta \geq 1$. If the initial approximation $\mathbf{x}^0 = c^0\mathbf{v}_1 + s^0\mathbf{u}^0$ is such that

$$|s^0|^{\alpha+\beta-1} \;<\; \frac{|\lambda_2 - \lambda_1|}{2C_1(1+C_2)}(1 - C_3), \tag{3.36}$$

then $t^i \to 0$, hence $\mathbf{x}^i \to \mathbf{v}_1$ and $\varrho(\mathbf{x}^i) \to \lambda_1$.

Proof: We carry on with writing τ^i_{ICM} for the residual condition in Inverse Correction and $\widetilde{\tau}^i$ for the one in inexact inverse iteration. Given $\mathbf{x}^i$, σ^i and τ^i_{ICM} we define

$$\Omega^i_{ICM} := \Omega^i_{ICM}(\mathbf{x}^i, \sigma^i, \tau^i_{ICM}) := \Big\{\mathbf{x}^{i+1} | \exists \mathbf{z}^i \text{ such that } \mathbf{x}^{i+1} = \frac{\mathbf{x}^i - \mathbf{z}^i}{\|\,\mathbf{x}^i - \mathbf{z}^i\,\|_2} \text{ and } \|\mathbf{r}^i - (A - \sigma^i I)\mathbf{z}^i\|_2 \leq \tau^i_{ICM}\Big\}. \tag{3.37}$$

Similarly we define for inexact inverse iteration

$$\widetilde{\Omega}^i := \widetilde{\Omega}^i(\mathbf{x}^i, \sigma^i, \widetilde{\tau}^i) := \Big\{\mathbf{x}^{i+1} | \exists \mathbf{y}^i \text{ such that } \mathbf{x}^{i+1} = \frac{\mathbf{y}^i}{\|\,\mathbf{y}^i\,\|_2} \text{ and } \|\mathbf{x}^i - (A - \sigma^i I)\mathbf{y}^i\|_2 \leq \widetilde{\tau}^i\Big\}. \tag{3.38}$$

The key idea of this proof is to show that $\Omega^i_{ICM} \subset \widetilde{\Omega}^i$ if $\widetilde{\tau}^i := |\varrho^i - \sigma^i|^{-1}\,\tau^i_{ICM}$.

Without loss of generality we consider the case $\varrho^i > \sigma^i$. For $\mathbf{x}^{i+1}_{ICM} \in \Omega^i_{ICM}$ there exists $\mathbf{z}^i$ such that $\mathbf{res}^i_{ICM} := \mathbf{r}^i - (A - \sigma^i I)\mathbf{z}^i$ has $\|\mathbf{res}^i_{ICM}\|_2 \leq \tau^i_{ICM}$ and $\mathbf{x}^i_{ICM} = \|\mathbf{x}^i - \mathbf{z}^i\|_2^{-1}\,(\mathbf{x}^i - \mathbf{z}^i)$. Now set $\widetilde{\mathbf{y}^i} = |\varrho^i - \sigma^i|^{-1}\,(\mathbf{z}^i - \mathbf{x}^i)$. Hence

$$\begin{aligned}
\|\mathbf{x}^i - (A - \sigma^i I)\widetilde{\mathbf{y}}^i\|_2 &= |\varrho^i - \sigma^i|^{-1}\|(\sigma^i - \varrho^i)\mathbf{x}^i - (A - \sigma^i)(\mathbf{z}^i - \mathbf{x}^i)\|_2 \\
&= |\varrho^i - \sigma^i|^{-1}\|A\mathbf{x}^i - \varrho^i\mathbf{x}^i - (A - \sigma^i)\mathbf{z}^i\|_2 \\
&= |\varrho^i - \sigma^i|^{-1}\|\mathbf{r}^i - (A - \sigma^i)\mathbf{z}^i\|_2 \\
&\leq |\varrho^i - \sigma^i|^{-1}\,\tau^i_{ICM} \;\leq\; \widetilde{\tau}^i,
\end{aligned}$$

hence $\widetilde{\mathbf{x}}^{i+1} = \|\widetilde{\mathbf{y}}^i\|_2^{-1}\,\widetilde{\mathbf{y}}^i \in \widetilde{\Omega}^i$. Finally we observe that the conditions leading to (3.34)

are satisfied, hence $\widetilde{\mathbf{x}}^{i+1} = \text{sign}(\varrho^i - \sigma^i)\mathbf{x}_{ICM}^{i+1}$ and so $\mathbf{x}_{ICM}^{i+1} \in \widetilde{\Omega}^i$. Therefore we can apply the convergence result for inexact inverse iteration Theorem 2.1 with $\widetilde{\tau}^i$. As the conditions of Theorem 2.1 are satisfied the claimed convergence follows immediately. □

In the case of $\sigma^i = \varrho(\mathbf{x}^i)$, which is excluded in Lemma 3.14, $\mathbf{z}^i = \mathbf{x}^i$ is the exact solution. Using $\mathbf{z}^i = \mathbf{x}^i$ leads to $\mathbf{y}^i = \mathbf{0}$, which is useless as it contains no information about the sought eigenvector. For example the Jacobi-Davidson method uses $\sigma^i = \varrho(\mathbf{x}^i)$, but uses a projected shifted matrix of the form

$$(I - \mathbf{x}^i(\mathbf{x}^i)^T)(A - \varrho(\mathbf{x}^i)I)(I - \mathbf{x}^i(\mathbf{x}^i)^T),$$

and therefore prohibits $\mathbf{z}^i = \mathbf{x}^i$.

For practical use the additional condition $\| \mathbf{res}^i \| \leq \frac{1}{2} \| \mathbf{r}^i \|$ is sensible. This condition is implied by those in Lemma 3.14, but can be missed if an estimator is used for $|s^i|$. To understand the need for doing this assume $\| \mathbf{res}^i \| \geq \| \mathbf{r}^i \|$ is permitted. Then $\mathbf{z}^i = \mathbf{0}$ is a valid approximation for the linear system $(A - \sigma^i I)\mathbf{z}^i = \mathbf{r}^i$ and hence with $\mathbf{x}^{i+1} = \mathbf{x}^i$ the iteration stagnates. Due to this argument we use

$$\tau_{ICM}^i \leq \min\left\{\frac{1}{2}\|\mathbf{r}^i\|_2, C_{10}\,|\varrho^i - \sigma^i|\left(\frac{\|\mathbf{r}^i\|_2}{|\varrho^i|}\right)^\beta, |\varrho^i - \sigma^i|\,C_3\,|c^i|\right\} \tag{3.39}$$

for some $C_{10} \in \mathbb{R}^+$ instead of (3.35).

For $|\lambda_1 - \sigma^i| \leq C_1\,|s^i|^\alpha$ with $\alpha > 0$ we get $|\varrho^i - \sigma^i| \leq \widetilde{C}_1\,|s^i|^\delta$, for some $\widetilde{C}_1 > 0$ and $\delta = \min\{2, \alpha\}$. Using the residual condition (3.39) we obtain for $\beta = 1$ that the residual needs to be of order $O(|s^i|^{\delta+1})$ as

$$\|\mathbf{res}^i\| \leq |\varrho^i - \sigma^i|\,C_{10}\,\|\mathbf{r}^i\|\,|\varrho|^{-1} = O(|s^i|^{\delta+1}).$$

Hence the initial advantage of an easier system to be solved vanishes when $\alpha > 0$ as the system gets more singular and simultaneously a more accurate solution needs to be obtained.

In Section 3.8 we provide some numerical tests illustrating the convergence. We tested the Inverse Correction method with fixed shift and various choices for variable shifts. The Inverse Correction Method is robust but slow and inefficient, contrarily the other variations are promisingly efficient but so far not robust.

Efficiency Analysis

While the convergence of the Inverse Correction Method followed from the convergence of inexact inverse iteration, this is no longer the case for the efficiency analysis. However we can use the same link to obtain the one step bound (2.18) from which the analysis runs similar to Theorem 3.10. In the following we provide an efficiency result similar to

previous results in Sections 3.3, 3.4 and 3.6.1. However this result is not as satisfactory as the earlier ones as in our experience it does not reflect the practical behaviour.

Theorem 3.15 *Consider that the inverse correction method, see Algorithm 3 is applied to find the eigenpair* $(\lambda_1, \mathbf{v}_1)$ *of* $A \in \mathbb{R}^{n\times n}$, *symmetric. Assume the conditions of Lemma 3.14 are satisfied. Additionally assume that* $\exists C_7 \in \mathbb{R}^+$ *such that*

$$|\varrho^i - \sigma^i||s^i| \;\leq\; C_7 \,\|\mathbf{res}^i\|_2, \tag{3.40}$$

then $\forall\ \Gamma \subset \mathbb{N}_n$ *with* $1 \in \Gamma$ *there exists* p_Γ, $q_\Gamma \in \mathbb{R}^+$ *with* $q_\Gamma < 1$ *independent of* σ^i *and* $\|\mathbf{res}^i\|_2$ *such that the number of MINRES iterations* $\mathcal{L}^i$ *in each outer iteration is bounded as*

$$\mathcal{L}^i \;\leq\; 1+ |\Gamma| + \frac{\log \dfrac{|t^i|}{|t^{i+1}|}}{\log {q_\Gamma}^{-1}} + C_{11} + \frac{\log \dfrac{1}{|\varrho^i - \sigma^i|}}{\log {q_\Gamma}^{-1}} \tag{3.41}$$

and the total number of MINRES iterations as

$$\mathcal{T} \;\leq\; \frac{\gamma \log 10 + \log \dfrac{|t^0|\ 10^{-\gamma}}{|t^N|}}{\log {q_\Gamma}^{-1}} + N(1+ |\Gamma| + C_{11}) + \sum_{i=0}^{N-1} \frac{\log \dfrac{1}{|\varrho^i - \sigma^i|}}{\log {q_\Gamma}^{-1}}, \tag{3.42}$$

where

$$C_{11} := \log\left(\frac{p_\Gamma \,\|P_1^{-1}\|}{|\lambda_2 - \lambda_1|} 8\,|\lambda_n - \lambda_1|\right) / \log((q_\Gamma)^{-1}).$$

Proof: As the conditions of Lemma 3.14 are satisfied we use again the relation between the Inverse Correction Method and inexact inverse iteration. Again we use the tilde to indicate variables of inexact inverse iteration and the subindex $_{ICM}$ for those of Inverse Correction. Hence we gain

$$\begin{aligned} t^{i+1} \;&\leq\; \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}\, \frac{|s^i| + \|\widetilde{\mathbf{res}^i}\|_2}{|c^i| - \|\widetilde{\mathbf{res}^i}\|_2} \\ &\leq\; \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|}\, \frac{|s^i| + \dfrac{\|\mathbf{res}^i_{ICM}\|_2}{|\varrho^i - \sigma^i|}}{|c^i| - \dfrac{\|\mathbf{res}^i_{ICM}\|_2}{|\varrho^i - \sigma^i|}}. \end{aligned}$$

Using the additional condition, (3.40), together with the conditions of Lemma 3.14 we gain

$$t^{i+1} \;\leq\; 8\frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \lambda_1|}(1 + C_8)\frac{\tau^i_{ICM}}{|\varrho^i - \sigma^i|}.$$

Rearranging this we gain

$$(\tau_{ICM}^i)^{-1} \, |\lambda_1 - \sigma^i|^{-1} \;\leq\; \frac{8(1+C_8)}{t^{i+1}\,|\lambda_2 - \lambda_1|} \, |\varrho^i - \sigma^i|^{-1} \, . \tag{3.43}$$

Now we apply the convergence result for preconditioned MINRES, Corollary 3.7, with $\mathbf{b}^i = \mathbf{r}^i$ to gain that $\exists\; p_\Gamma$ and q_Γ such that $\forall\; \Gamma \subset \mathbb{N}_n$ with $1 \in \Gamma$ and for all $k \in \mathbb{N}$

$$\|\mathbf{res}_k^i\|_{P^{-1}} \;\leq\; (q_\Gamma)^{k-|\Gamma|} p_\Gamma \, |\lambda_1 - \sigma^i|^{-1} \|Q_\Gamma (W^i)^T P_1^{-1} \mathbf{x}^i\|_2 \, .$$

If preconditioned MINRES is used, then $P_1^2 = P$ is the preconditioner, otherwise $P_1 = I$. We can bound $\|Q_\Gamma (W^i)^T P_1^{-1} \mathbf{x}^i\|_2 \leq \|P_1^{-1}\| \|\mathbf{r}^i\|$. Using the definition for $\mathcal{L}^i$, Definition 3.8, we have

$$\tau_{ICM}^i \;\leq \|\mathbf{res}_{\mathcal{L}^i-1}^i\|_{P_1^{-1}} \leq\; (q_\Gamma)^{\mathcal{L}^i - 1 - |\Gamma|} p_\Gamma \, |\lambda_1 - \sigma^i|^{-1} \|P_1^{-1}\|_2 \|\mathbf{r}^i\|_2 \, .$$

Rearranging we gain

$$\mathcal{L}^i \;\leq\; 1 + |\Gamma| + \log\left(\frac{p_\Gamma \, \|P_1^{-1}\| \|\mathbf{r}^i\|}{\tau_{ICM}^i \, |\lambda_1 - \sigma^i|} \right) / \log((q_\Gamma)^{-1}). \tag{3.44}$$

Next we insert (3.43) into (3.44) to gain

$$\begin{aligned}
\mathcal{L}^i \;&\leq\; 1 + |\Gamma| + \log\left(\frac{p_\Gamma \, \|P_1^{-1}\| \|\mathbf{r}^i\| \, 8}{|\lambda_2 - \lambda_1| |\varrho^i - \sigma^i| \, t^{i+1}} \right) / \log((q_\Gamma)^{-1}) \\
&\leq\; 1 + |\Gamma| + \log\left(\frac{p_\Gamma \, \|P_1^{-1}\| |\lambda_n - \lambda_1|}{|\lambda_2 - \lambda_1| |\varrho^i - \sigma^i| \, t^{i+1}} \right) / \log((q_\Gamma)^{-1}) \\
&\quad + \log \frac{t^i}{t^{i+1}} / \log((q_\Gamma)^{-1}) + \log(|\varrho^i - \sigma^i|^{-1}) / \log((q_\Gamma)^{-1}).
\end{aligned}$$

Substituting C_{11} we gain (3.41) and further by summing over $\mathcal{L}^i$ for $i = 0, \ldots, \mathcal{N} - 1$ we gain (3.42). □

The bounds given in Lemma 3.15 are valid for unpreconditioned MINRES as well as preconditioned MINRES, just the values of p_Γ and q_Γ differ. Comparing Lemma 3.15 with the corresponding bound for $\mathcal{L}^i$ for inexact inverse iteration using unpreconditioned MINRES, Theorem 3.10, we observe an additional term in Theorem 3.15. However this additional term does not agree fully with our practical experience, see Example 3.6 in Section 3.8. Further we experienced that convergence is either slow, this is the case for a fixed shift, or erratic and non robust for variable shifts. For corresponding result see Section 3.8 Example 3.6 and Table 3.9 for fixed shift and Tables 3.10, 3.11 and 3.12 for variable shifts.

Finally we comment on the algorithm proposed by Golub and Ye (2000), geared towards the generalised eigenvalue problem. The first iteration is a standard step of

inverse iteration using an inexact solution of a shifted linear system. In the remaining outer iterations a scaled eigenvalue residual $\mathbf{r}_{GY}^i = -\mathbf{r}^i\phi^i$ is computed, where $\phi^i := \|\mathbf{y}^{i-1}\|_\infty$. This residual is then taken as the right hand side for the standard system $A - \sigma I$ and the solution is added to the old solution. By dividing the scaled residual and the solution of the linear system by ϕ we obtain the Inverse Correction Method with fixed shift σ^i. In Section 4.3.5 we discuss the method from Golub and Ye (2000) in more detail, further we also consider their stopping condition.

3.6 Preconditioned Inexact Inverse Iteration with MINRES (PInvit)

In the previous sections we discussed two approaches of preconditioning inexact inverse iteration using MINRES. However for a variety of reasons none of them was completely satisfactory. The preconditioning of the standard equation, studied in Section 3.4, leads to a bound on the total cost which includes the term $\sum \log(1/t^{i+1})$, which has the potential of creating large additional costs. Slow convergence or loss of robustness is the downfall of the approach by Rüde and Schmid (1995). Here we discuss the approach from Simoncini and Eldén (2002) which is based on the observation from Scott (1981). We extend this approach to arbitrary but fixed positive definite preconditioners.

A brief discussion of key ideas of Simoncini and Eldén (2002) is given in Section 3.6.1. In Section 3.6.2 we state and analyse the convergence of the algorithm PInvit, which is a generalisation of inexact inverse iteration in the sense that we allow general right-hand sides $\mathbf{b}^i$. The convergence analysis is again independent of the applied linear solver and extends the Theorem 2.1 to modified right-hand sides. For the remaining sections we then consider MINRES as linear solver and discuss the choice of the right hand side in Section 3.6.3. The key result will be based on the observation that modifying the right-hand side in the way Simoncini and Eldén (2002) suggest reduces the cost of a linear solve. Then is Section 3.6.4 we state the efficiency result, and conclude the discussion on PInvit by explaining a few implementational aspects of MINRES with regard to the modified right hand side. In Section 3.6.5 we discuss how the right-hand side can be obtained if neither the action of P_1 nor P on a vector is available.

3.6.1 Approach by Simoncini and Elden

Simoncini and Eldén (2002) consider the combination of inexact inverse iteration and Galerkin-Krylov techniques, namely GMRES, MINRES and CR. Their analysis ties the Krylov solver with the outer iteration. By doing so they derive in the unpreconditioned case a convergence result based on the reduction of the eigenvalue residual $\|\mathbf{r}^i\|$. Further they suggest a new stopping condition for the linear solves, which we discuss later, (3.47). Another key result, motivated by Scott (1981), is to consider, in

the case of Cholesky preconditioned MINRES solves, an alternative system to update the eigenvector approximation. Instead of the standard preconditioned system

$$P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{z}^i = P_1^{-1}\mathbf{x}^i \quad \text{with} \quad \mathbf{y}^i = P_1^{-T}\mathbf{z}^i, \tag{3.45}$$

they consider for the same system matrix a modified right hand side

$$P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{z}^i = P_1^{T}\mathbf{x}^i \quad \text{with} \quad \mathbf{y}^i = P_1^{-T}\mathbf{z}^i. \tag{3.46}$$

Here $P_1P_1^T$ is an incomplete Cholesky factorisation of A, assuming A is symmetric positive definite. In case of unpreconditioned solves, $P_1 = I$, both systems are equivalent and the right hand sides is an approximation of the sought eigenvector. The modification of Simoncini and Eldén (2002) preserves this quality for the preconditioned system, which, as we see later, reduces the cost of the linear solver.

To link the inner iterations with the outer iteration they use a stopping condition for MINRES based on the reduction of the eigenvalue residual. As a stopping condition for the inner iteration they use

$$\frac{|\ \|\mathbf{y}_k^i\| - \|\mathbf{y}_{k-1}^i\|\ |}{\|\mathbf{y}_k^i\|} \leq \tau_{SE} \quad \text{and} \quad \|\mathbf{y}_k^i\| \geq \frac{1}{\|\mathbf{r}^i\|}, \tag{3.47}$$

where both conditions need to be satisfied simultaneously in order to stop the inner iteration. The second condition ensures that the eigenvalue residual is reduced. However the first condition, inspired by their convergence analysis which uses the fact that P_1 is a Cholesky preconditioner, is a heuristic to stop MINRES when the progress achieved with respect to the outer process starts to deteriorate. Later in Section 3.8 we compare this combined stopping condition from Simoncini and Eldén (2002), (3.47), against the standard residual condition $\|\mathbf{res}\| \leq \tau$, see Test 3.5.

Here we analyse the approach from Simoncini and Eldén (2002) in a more general setting in order to extend it to other preconditioners.

3.6.2 Convergence

In this section we consider instead of the standard linear system $(A-\sigma^i I)\mathbf{y}^i = \mathbf{x}^i$, the modified system

$$(A-\sigma^i I)\mathbf{y}^i = \mathbf{b}^i, \tag{3.48}$$

which, when symmetric preconditioning is applied with $P = P_1P_1^T$, has the form

$$P_1^{-1}(A-\sigma^i I)P_1^{-T}\widetilde{\mathbf{y}}^i = P_1^{-1}\mathbf{b}^i, \quad \text{with} \quad \mathbf{y}^i = P_1^{-T}\widetilde{\mathbf{y}}^i. \tag{3.49}$$

Algorithm 4: Generalised Inexact Inverse Iteration

- Given $\mathbf{x}^0$, and $C_3 > 0$,
- For $i = 0, 1, 2, 3, \ldots$
  - Choose $\mathbf{b}^i$ such that $|\mathbf{v}_1^T \mathbf{b}^i| \geq C_3$, further choose σ^i and τ^i,
  - Solve $(A - \sigma^i I)\mathbf{y}^i = \mathbf{b}^i$ such that $\|\mathbf{b}^i - (A - \sigma^i I)\mathbf{y}^i\| \leq \tau^i$,
  - Update $\mathbf{x}^{i+1} = \mathbf{y}^i / \|\mathbf{y}^i\|$,
  - Test for convergence

First we derive a one step bound similar to the one for the standard system (2.18). In order to derive this we define the residual similar to (2.10),

$$\mathbf{res}^i := \mathbf{b}^i - (A - \sigma^i I)\mathbf{y}^i \tag{3.50}$$

and use the same orthogonal splitting as in Chapter 2,

$$\mathbf{x}^i = c^i \mathbf{v}_1 + s^i \mathbf{u}^i, \tag{3.51}$$

with $\mathbf{v}_1$, $\mathbf{u}^i$, and $\mathbf{p}^i$ orthonormal. We start with rearranging the residual equation (3.50)

$$(A - \sigma^i I)\mathbf{y}^i = \mathbf{b}^i + \mathbf{res}^i. \tag{3.52}$$

Now we premultiply by $\mathbf{v}_1^T (A - \sigma^i I)^{-1}$ while assuming $0 < |\lambda_1 - \sigma^i| \leq \frac{1}{2} |\lambda_2 - \lambda_1|$ to obtain the cosine equation

$$\|\mathbf{y}^i\| \, c^i = (\lambda_1 - \sigma^i)^{-1} (\mathbf{v}_1^T \mathbf{b}^i + \mathbf{v}_1^T \mathbf{res}^i).$$

From which we obtain a lower bound on $|c^i|$ by assuming $|\mathbf{v}_1^T \mathbf{b}^i| > \|\mathbf{res}^i\|_2$,

$$\|\mathbf{y}^i\| |c^i| \geq |\lambda_1 - \sigma^i|^{-1} \, (|\mathbf{v}_1^T \mathbf{b}^i| - \|\mathbf{res}^i\|_2). \tag{3.53}$$

Next we premultiply (3.52) by $(I - \mathbf{v}_1 \mathbf{v}_1^T)(A - \sigma^i I)^{-1}$ to gain the sine equation

$$\|\mathbf{y}^i\| \, s^{i+1} \mathbf{u}^{i+1} = (I - \mathbf{v}_1 \mathbf{v}_1^T)(A - \sigma^i I)^{-1} (\mathbf{b}^i + \mathbf{res}^i),$$

which by taking norms gives

$$\|\mathbf{y}^i\||s^{i+1}| \leq |\lambda_2 - \sigma^i|^{-1} \left(\|(I - \mathbf{v}_1\mathbf{v}_1^T)\mathbf{b}^i\| + \|\mathbf{res}^i\|\right). \tag{3.54}$$

Finally we divide the sine inequality (3.54) by the cosine inequality (3.53) to obtain the one step bound

$$t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \frac{\|(I - \mathbf{v}_1\mathbf{v}_1^T)\mathbf{b}^i\| + \|\mathbf{res}^i\|}{|\mathbf{v}_1^T\mathbf{b}^i| - \|\mathbf{res}^i\|}. \tag{3.55}$$

We only require conditions

$$|\mathbf{v}_1^T\mathbf{b}^i| \geq C_3 \tag{3.56}$$

and $\|\mathbf{res}^i\| \leq C_2$ for some suitable constants C_2 and C_3 to prove convergence based on the shift tending towards the desired eigenvalue.

Lemma 3.16 *Consider Algorithm 4 being applied to $A \in \mathbb{R}^{n\times n}$ symmetric. Assume $\exists$ C_1, $C_3 \in \mathbb{R}^+$ and $C_2 \in (0,1)$ while $\alpha \geq 1$ such that the shift σ^i satisfies*

$$0 < |\lambda_1 - \sigma^i| < \min\{C_1\,|s^i|^\alpha, \frac{1}{2}\,|\lambda_2 - \lambda_1|\}$$

and that the residual condition τ^i satisfies $\tau^i \leq C_2\,|\mathbf{v}_1^T\mathbf{b}^i|$ while the right-hand side satisfies $C_3\,|\mathbf{b}^i| \leq |\mathbf{v}_1^T\mathbf{b}^i|$. If the initial approximation $\mathbf{x}^0 = c^0\mathbf{v}_1 + s^0\mathbf{w}^0$ is such that

$$|s^0| < \frac{(1 - C_2)C_3}{2(1 + C_2)C_1}\,|\lambda_2 - \lambda_1|$$

then $t^i \to 0$ and $\varrho^i \to \lambda_1$ while $\mathbf{x}^i \to \mathbf{v}_1$.

Proof: Starting with the one-step bound (3.55) we have

$$\begin{aligned} t^{i+1} &\leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \frac{\|(I - \mathbf{v}_1\mathbf{v}_1^T)\mathbf{b}^i\| + \|\mathbf{res}^i\|}{|\mathbf{v}_1^T\mathbf{b}^i| - \|\mathbf{res}^i\|} \\ &\leq t^i\,|s^i|^{\alpha-1}\,\frac{2C_1}{|\lambda_2 - \lambda_1|}\,\frac{(1 + C_2)\,\|\mathbf{b}^i\|}{(1 - C_2)\,|\mathbf{v}_1^T\mathbf{b}^i|} \\ &\leq t^i\,|s^i|^{\alpha-1}\,\frac{2C_1}{|\lambda_2 - \lambda_1|}\,\frac{(1 + C_2)}{(1 - C_2)C_3} \end{aligned}$$

With $C_4 := 2C_1\,|s^i|^{\alpha-1}\,(1+C_2)(1-C_2)^{-1}C_3^{-1}\,|\lambda_2 - \lambda_1|^{-1} < 1$ we gain $t^i \leq (C_4)^i t^0 \to 0$ and therefore $\varrho^i \to \lambda_1$ and $\mathbf{x}^i \to \mathbf{v}_1$. □

Let $\mathbf{b}^i = \xi_1\mathbf{x}^i + \xi_2\widetilde{\mathbf{b}}^i$ for some $\widetilde{\mathbf{b}}^i \in \mathbb{R}^n$ orthonormal to $\mathbf{v}_1$, then

$$\|(I - \mathbf{v}_1\mathbf{v}_1^T)\mathbf{b}^i\|_2 \leq |\xi_1||s^i| + |\xi_2|\,.$$

If additionally $\|\mathbf{res}^i\| = O(|s^i|)$ then

$$\frac{\|(I - \mathbf{v}_1\mathbf{v}_1^T)\mathbf{b}^i\|_2 + \|\mathbf{res}^i\|_2}{|\mathbf{v}_1^T\mathbf{b}^i| - \|\mathbf{res}^i\|_2} = O(|s^i|) + O(|\xi_2|).$$

Hence for the outer convergence the choice $\xi_2 = 0$ is beneficial. This choice for $\mathbf{b}^i$ leads to inexact inverse iteration using preconditioned MINRES, as discussed in Section 3.4. Here we look for some alternative choice for $\mathbf{b}^i$ to improve the overall performance. By discarding the choice of $\mathbf{b}^i$ which is beneficial for the outer convergence we are free to find a $\mathbf{b}^i$ reducing the cost of a linear solve.

3.6.3 Right-hand side $\mathbf{b}^i$

From now on we use the fact that the linear solver is MINRES. In Section 3.2 we already provided with Corollary 3.7 a suitable convergence result for preconditioned MINRES. The residual bound given there is for the system (3.49). However when ν_1 denotes the largest singular value of P_1, i.e. the square root of the largest eigenvalue of P, then we get with Corollary 3.7

$$\begin{aligned} \|\mathbf{res}_k^i\| &\leq \nu_1 \|P_1^{-1}\mathbf{res}_k^i\| \\ &\leq \nu_1 \|P_1^{-1}\mathbf{b}^i - P_1^{-1}(A - \sigma I)P_1^{-T}\tilde{\mathbf{y}}_k^i\| \\ &\leq \nu_1 (q_\Gamma)^{k-|\Gamma|} p_\Gamma \, |\lambda_1 - \sigma^i|^{-1} \, \chi^i, \end{aligned} \tag{3.57}$$

where p_Γ and q_Γ are as defined in Lemma 3.5. Here $\chi^i = \|Q_\Gamma (W^i)^T P_1^{-1}\mathbf{b}^i\|_2$ and W^i the matrix of eigenvectors of the preconditioned system matrix, $P_1^{-1}(A - \sigma I)P_1^{-T}$. We recall that the matrix Q_Γ is the projection matrix defined in (3.11). However the factor P_1 is only needed in theory, its existence is ensured as P is symmetric positive definite.

In order to derive an appropriate bound on χ^i we use a perturbation result from Chatelin (1993).

Lemma 3.17 *Given $B \in \mathbb{R}^{n\times n}$ symmetric with eigen-decomposition $B = W\Lambda_B W^T$ and simple eigenpair $(\mu_1, \mathbf{w}_1)$ then $B + \Delta B$ has an eigenpair $(\mu_1 + \Delta\mu_1, \mathbf{w}_1 + \Delta\mathbf{w}_1)$ with $\Delta\mu_1 = \mathbf{w}_1^T \Delta B \mathbf{w}_1 + O(\|\Delta B\|^2)$ and*

$$\tan\angle(\mathbf{w}_1, \mathbf{w}_1 + \Delta\mathbf{w}_1) = \|(\Lambda_B - \mu_1 I)^{-D} W^T \Delta B \mathbf{w}_1\|_2 + O(\|\Delta B\|^2),$$

where $(\Lambda_B - \mu_1 I)^{-D}$ is the Drazin inverse of $\Lambda_B - \mu_1 I$.

Proof: See Chatelin (1993, Proposition 4.2.1 and corresponding proof). □

As $\Lambda_B - \mu_1 I = \mathrm{diag}(0, (\mu_1 - \mu_1), \ldots, (\mu_n - \mu_1))$ the Drazin inverse is given by $(\Lambda_B - \mu_1 I)^{-D} = \mathrm{diag}(0, (\mu_1 - \mu_1)^{-1}, \ldots, ^{-1}, (\mu_n - \mu_1)^{-1})$.

Based on Lemma 3.17 we can derive a bound for $\chi^i = \|Q_\Gamma (W^i)^T P_1^T \mathbf{x}^i\|_2$.

Corollary 3.18 *Let $(\sigma^i, \mathbf{x}^i)$ be an approximation to the eigenpair $(\lambda_1, \mathbf{v}_1)$ of $A \in \mathbb{R}^{n\times n}$ symmetric, where λ_1 is simple and let $(\mu_1^i, \mathbf{w}_1^i)$ be the eigenpair of $P_1^{-1}(A-\sigma^i I)P_1^{-T}$ closest to zero. If $|s^i| = \sin\angle(\mathbf{x}^i, \mathbf{v}_1)$ for small enough $|s^i|$ and $|\sigma^i - \lambda_1| \leq C_1\,|s^i|$, then $\exists C_{10} \in \mathbb{R}^+$ such that*

$$|\sin\angle(P_1^T\mathbf{x}^i, \mathbf{w}_1^i)| \;\leq\; C_{10}\,|s^i|\,.$$

Proof: Let $\mathbf{w}_1$ be the eigenvector corresponding to the simple eigenvalue zero of $P_1^{-1}(A-\lambda_1 I)P_1^{-T}$. Then we make use of

$$\begin{aligned}&|\sin\angle(P_1^T\mathbf{x}^i, \mathbf{w}_1^i)|\\ \leq\; &|\sin\angle(P_1^T\mathbf{x}^i, P_1^T\mathbf{v}_1)| + |\sin\angle(P_1^T\mathbf{v}_1, \mathbf{w}_1)| + |\sin\angle(\mathbf{w}_1, \mathbf{w}_1^i)|\,.\end{aligned}$$

Let $\mathbf{x}^i = c^i\mathbf{v}_1 + s^i\mathbf{u}^i$ with $\|\,\mathbf{v}_1\,\|_2 = \|\,\mathbf{u}^i\,\|_2 = 1$ and $\mathbf{u}^i \perp \mathbf{v}_1$, then as P_1 non singular $|\sin\angle(P_1^T\mathbf{x}^i, P_1^T\mathbf{v}_1)| \leq \nu_n\nu_1^{-1}\,|s^i|$, where $\nu_1 = \nu_1(P_1)$ is the smallest singular value of P_1 and $\nu_n = \nu_n(P_1)$ the largest singular value of P_1. Further as $P_1^{-1}(A-\lambda_1 I)P_1^{-T}(P_1^T\mathbf{v}_1) = 0$ we gain $\sin\angle(P_1^T\mathbf{v}_1, \mathbf{w}_1) = 0$. Finally as $|\lambda_1 - \sigma^i| \leq C_1\,|s^i|$ we gain by using Lemma 3.17 that for some $\widetilde{C}_6 \in \mathbb{R}^+$, $|\sin\angle(\mathbf{w}_1, \mathbf{w}_1^i)| \leq \widetilde{C}_6\,|s^i|$. □

As $\left(P_1^{-1}(A-\lambda_1 I)P_1^{-T}\right)P_1^{-1}(P\mathbf{v}_1) = 0$, we have with $P_1^{-1}P\mathbf{v}_1 = P_1^T\mathbf{v}_1$ an eigenvector of $P_1^{-1}(A-\lambda_1 I)P_1^{-T}$ corresponding to the eigenvalue 0. Let as usual $|\lambda_1 - \sigma^i| < \frac{1}{2}\,|\,\lambda_2 - \lambda_1\,|$ and then we gain from Corollary 3.18 that $\exists C_{10} > 0$ such that for $|\lambda_1 - \sigma^i| < \frac{1}{2}\,|\lambda_2 - \lambda_1|$

$$\|Q_\Gamma(W^i)^T P_1^{-1}\mathbf{b}^i\|_2 \;\leq\; |s^i|\|P_1\|\,C_{10}.$$

Remark 3.19 *Therefore $\mathbf{b}^i = P\mathbf{x}^i$ is a good right-hand side in the sense that it reduces the cost of a linear solve.*

In the following we show how preconditioned MINRES can be adapted to provide $\mathbf{z}^i = P\mathbf{y}^i$.

3.6.4 Efficiency

The key result is again similar to Theorem 3.10.

Theorem 3.20 *Assume the conditions of Lemma 3.16 being satisfied then PInvit using MINRES converges. Let P denote the preconditioner and choose $\mathbf{b}^i = P\mathbf{x}^i$ in Algorithm 4. If there exists $C_5 \in R^+$ such that the residual satisfies $|s^i| \leq C_5\,\|\mathbf{res}^i\|$, then the number of preconditioned MINRES iterations $\mathcal{L}^i$, see Definition 3.8, satisfies*

$$\mathcal{L}^i \;\leq\; 1 + |\Gamma| + \log(C_9 C_{10}\frac{t^i}{t^{i+1}})/\log((q_\Gamma)^{-1}), \tag{3.58}$$

with

$$C_9 \;:=\; \frac{2\nu_1(1+C_5)p_\Gamma}{|\lambda_2-\lambda_1|\,(1-C_2)C_3}.$$

Further if $\mathbf{x}^0$ *is to be improved by a factor* $10^{-\gamma}$ *then the total number of MINRES iteration is bounded by*

$$\mathcal{T} \;\le\; \frac{\gamma\log 10+\log\dfrac{10^{-\gamma}t^0}{t^{\mathcal{N}}}}{\log((q_\Gamma)^{-1})}+\mathcal{N}\left(1+|\Gamma|+\frac{\log(C_{10}C_9)}{\log((q_\Gamma)^{-1})}\right). \tag{3.59}$$

Proof: As the conditions of Corollary 3.7 are satisfied we obtain with (3.57) the following bound for the residual

$$\|\mathbf{res}_k^i\| \;\le\; \nu_1(q_\Gamma)^{k-|\Gamma|}p_\Gamma\,|\lambda_1-\sigma^i|^{-1}\,\chi^i.$$

For each i, $|\lambda_1-\sigma^i|>0$ and χ^i are fixed and $q_\Gamma\in(0,1)$ therefore $\|\mathbf{res}_k^i\|\to 0$. Hence for all i there exists $\mathcal{L}^i$ as defined in Definition 3.8. Applying Lemma 3.16 we obtain the claimed convergence.

To prove the bound on $\mathcal{L}^i$ we use

$$\tau^i \;<\|\mathbf{res}^i_{\mathcal{L}^i-1}\|_2\le\;\nu_1(q_\Gamma)^{\mathcal{L}^i-1-|\Gamma|}p_\Gamma\,|\lambda_1-\sigma^i|^{-1}\,\chi^i.$$

Rearranging and taking the logarithm gives

$$\mathcal{L}^i \;\le\; 1+|\Gamma|+\log\left(\frac{\nu_1 p_\Gamma\chi^i}{|\lambda_1-\sigma^i|\,\tau^i}\right)/\log((q_\Gamma)^{-1}). \tag{3.60}$$

Next we use the one step bound to gain

$$\begin{aligned}
t^{i+1} &\le \frac{|\lambda_1-\sigma^i|}{|\lambda_2-\sigma^i|}\frac{\|(I-\mathbf{v}_1\mathbf{v}_1^T)\mathbf{x}^i\|+\|\mathbf{res}^i\|}{|\mathbf{v}_1^T\mathbf{x}^i|-\|\mathbf{res}^i\|}\\
&\le 2\frac{|\lambda_1-\sigma^i|}{|\lambda_2-\lambda_1|}\frac{(C_5+1)\,\|\mathbf{res}^i\|}{(1-C_2)C_3}\\
&\le \frac{2(C_5+1)\,\|\mathbf{res}^i\|}{(1-C_2)C_3\,|\lambda_2-\lambda_1|}\tau^i\,|\lambda_1-\sigma^i|\,.
\end{aligned}$$

By rearranging we get

$$|\lambda_1-\sigma^i|^{-1}\,(\tau)^{-1} \;\le\; \frac{2(C_5+1)\,\|\mathbf{res}^i\|}{(1-C_2)C_3\,|\lambda_2-\lambda_1|\,t^{i+1}} \tag{3.61}$$

Inserting (3.61) into (3.60) we gain

$$\mathcal{L}^i \leq 1 + |\Gamma| + \log\left(\frac{2\nu_1(C_5+1)p_\Gamma\chi^i}{|\lambda_2-\lambda_1|\,(1-C_2)C_3t^{i+1}}\right)/\log((q_\Gamma)^{-1}). \tag{3.62}$$

In order to derive (3.58) we use Corollary (3.18) to gain for $i > 0$

$$\chi^i = \|Q_\Gamma(W^i)^T P_1^{-1}\mathbf{b}^i\| = \|Q_\Gamma(W^0)^T P_1^{-1}P\mathbf{x}^0\| = \|Q_\Gamma(W^0)^T P_1^T\mathbf{x}^0\| \leq C_{10}\,|s^i|\,.$$

Summing over $\mathcal{L}^i$ for $i = 0,1,2,\ldots,\mathcal{N}-1$ gives (3.59). □

The bound on $\mathcal{T}$, (3.59), has now the same form as for the unpreconditioned case, (3.22), while the constants p_Γ and q_Γ improve due to preconditioning. However the additional constant C_{10} is not quantified and in practise C_{10} might be large.

It is optimal to reduce the number of outer iteration which can be achieved by choosing the RQ as shift. In Chapter 2, we have seen that in practise the difference between quadratic convergence and cubic convergence in terms of the number of outer iterations $\mathcal{N}$ is negligible. Therefore it is better to use the right hand side to reduce the cost per outer iteration $\mathcal{L}^i$ than to improve on the number of outer iterations $\mathcal{N}$ (as long the quadratic convergence is preserved). Hence the use of the modified right hand side as in PInvit should reduce the cost in comparison to RQIf and RQId. This can be observed in practise, see Section 3.8 where we compare methods based on the modified right hand side against methods using the standard right hand side.

3.6.5 Adapted preconditioned MINRES

While for Cholesky preconditioning the modified right hand side $\mathbf{b} = P\mathbf{x}$ can easily be computed this is no longer the case for preconditioners based on Domain Decomposition or Multi Grid. Therefore we now study one way to provide this right hand side by using additional information from the previous linear solve. The treatment is specific for MINRES, but can be generalised to GMRES using left, right or centered preconditioning. We start by discussing a few implementational aspects of preconditioned MINRES. Throughout this section let P be the preconditioner, we only want to require the action of P^{-1} on a vector.

The MINRES algorithm, for example see Fischer (1996, p. 185), uses the Lanczos/Arnoldi sequence to construct a P-orthonormal basis U for the Krylov space $\mathcal{K}(P^{-1}(A-\sigma^i I), P^{-1}\mathbf{b}^i)$. That is U is constructed such that $P^{-1}(A-\sigma^i I)U = UT$ where T is tridiagonal (this follows from the fact that T is at least upper Hessenberg, $U^TPU = I$, and A symmetric). Then MINRES constructs a QR factorisation of T where Q is a sequence of Givens rotation and R is an upper right matrix. As T is tridiagonal R has all entries equal to zero except those on the diagonal and the first two upper diagonals. Based on this format of R there exists a three term recurrence

formula for R^{-1}. Let R be given by

$$R = \begin{pmatrix} r_{1,1} & r_{2,2} & r_{3,3} & & 0 \\ & r_{1,2} & r_{2,3} & \ddots & \\ & & r_{1,3} & \ddots & r_{3,n} \\ & & & \ddots & r_{2,n} \\ 0 & & & & r_{1,n} \end{pmatrix}.$$

Then the three term recurrence formula for R^{-1} is given by

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{e}_1/r_{1,1} \quad \mathbf{w}_2 = (\mathbf{e}_2 - r_{2,2}\mathbf{w}_1)/r_{1,2} \\ \mathbf{w}_k &= (\mathbf{e}_k - r_{2,k}\mathbf{w}_{k-1} - r_{3,k}\mathbf{w}_{k-2})/r_{1,k}, \end{aligned} \tag{3.63}$$

where $R^{-1} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$. This recurrence formula can be applied to any matrix B, such that the calculation of BR^{-1} needs only the same three term recurrence formula applied to the vectors of $B\mathbf{e}_k$, instead of the unit vectors $\mathbf{e}_k$. Finally the solution is constructed as $\mathbf{y} = UR^{-1}Q^T\mathbf{e}_1$. For a more general discussion of such recurrence formulas see Fischer (1996, Chapter 2).

To obtain $\mathbf{z} = P\mathbf{y}$ we only need to apply the three term recurrence formula (3.63) to $\widetilde{U} = PU$. To see how to obtain $\widetilde{U}$ we look how U is calculated. In the algorithm the kth vector of U is calculated by $U\mathbf{e}_k = P^{-1}(A - \sigma I)\mathbf{q}/\beta_k$ for some vector $\mathbf{q}$. Hence we can construct $\widetilde{U}\mathbf{e}_k = (A - \sigma I)\mathbf{q}/\beta_k$ without further applications of P or A. The calculation of $\mathbf{z}^i = P\mathbf{y}^i$ requires only four vectors storage and an additions of three vectors in each inner iteration additional to the cost of standard preconditioned MINRES. In a similar fashion $A\mathbf{x}$ can be provided without further matrix vector products or application of the preconditioner. This gives a cheap check for the residual condition. For a detailed algorithm see Appendix B.

3.7 Robustness and Stopping Conditions

As usual, due to effects of round off errors, there are differences between theory and practical experience. In this section we consider such differences and explore how we can gain a robust method.

Therefore we start with considering practical difficulties which might rise when using inexact inverse iteration with MINRES as a linear solver. Then we describe some counter-measures in form of additional stopping conditions.

3.7.1 Possible breakdowns and their source

Inconsistent systems

By our efficiency results in this chapter we have seen that letting the shift tend towards the desired eigenvalue is beneficial for the overall performance. However, since $A - \sigma I$ tends to a singular matrix, shifting towards the desired eigenvalue increases the danger of failure of the linear solver. We have studied empirically the effect of keeping a certain distance to the sought eigenvalue and the shift. For these tests we used shifts of the form $\sigma^i = \varrho^i + pert$ where $pert$ is a fixed perturbation. We did tests with $pert = 0$, $pert = 10eps\ \|A\|$ and $pert = 10^3 eps\ \|A\|$, where eps is the machine precision. There were no significant differences between these three choices of $pert$. Breakdowns occur with the same likelihood for all three choices.

More significant are the differences between the methods, meaning between unpreconditioned solves, and preconditioned solves with the standard and the modified right hand side. Using the modified right hand side improves the robustness considerably. This approach leads to the least number of breakdowns. In contrast, unpreconditioned solves only broke down when the subspace size was large, and round off errors prevented the detection of an acceptable solution. More interesting is that from all preconditioned approaches the combination of inexact inverse iteration and preconditioned MINRES as discussed in Section 3.4 failed most. Our explanation for this is that for the standard preconditioned system the first vector in the subspace is not an approximation of the solution. In contrast, in the unpreconditioned and also in the preconditioned case where the modified right hand side is used, the right hand side is an approximation of the sought solution. In the standard preconditioned case until the same approximation is regained the round off errors might destroy the chance of finding an accurate solution.

Summarising, we found that the almost singularity of the system is for our application not of great concern. The breakdowns were often triggered by some other reason, but possibly enforced by the singularity.

Round off errors

Another possible error source are round off errors made during the calculation of the Krylov basis and the projection of $A - \sigma^i I$ onto the Krylov subspace. The basis for the Krylov subspace should be orthogonal, however due to round off errors it will not be orthogonal. (The possible danger of losing the orthonormality of the Krylov basis somewhere during the process is closely linked with the convergence of the approximation to the exact solution.) Perhaps more significantly is that the improvements added to the current approximation gets inaccurate. As a result the estimator for the norm of the residual decreases while the norm of the residual itself may increase.

Too weak requirements

Obviously, if the tolerance and the shift are not tight enough, so not satisfying the conditions of the convergence results, then the resulting solution might not improve the previous approximation.

Too strong requirements

Consider $\mathbf{y} = B^{-1}\mathbf{b}$ and $\|\Delta\mathbf{y}\| \leq eps\ \|\mathbf{y}\|$, where *eps* is the machine precision, then $\mathbf{y} + \Delta\mathbf{y}$ are same solution and $\|B(\mathbf{y} + \Delta\mathbf{y}) - \mathbf{b}\| \leq eps\ \|B\|\|\mathbf{y}\|$. This shows that in practice the machine precision limits the achievable accuracy, both for the eigenvalue approximation and for the linear solve. (In both cases the convergence is in practice checked by an evaluation of a residual.) As we do not know how much accuracy is achievable for the eigenvalue problem, it might happen that we asked for a not achievable accuracy for the eigenpair approximation. As MINRES has to provide the outer iteration with the corresponding solution, this implies that MINRES does not find a solution of the specified accuracy. We now discuss the effect this has on MINRES.

For us the important feature is the departure between the internal estimator used within MINRES for the residual norm, $\|\mathbf{res}^i\|$, and a direct calculation of residual norm and its exact value (which not available). The direct calculation of the residual using the current approximation needs one matrix vector product, hence it is usual to use the internal estimator and calculate the residual only when the estimator is small enough. However as we use matrix vector products and possibly some preconditioner, the calculated residual might also differ considerably from the exact one. Further the estimator will differ from the exact one as the estimator does not stop its convergence towards zero when the norm of the exact residual stagnates.

Another feature of too strong requirements is described under the next header, round off level.

Round off level

When $\mathbf{x}$ is accurate up to round off level at least to the information of the algorithm, then no further improvement will be achieved. To make this more specific denote the j-th component of a vector $\mathbf{y}$ by $(\mathbf{y})_j$. Now if

$$|(\mathbf{y}_k)_j - (\mathbf{y}_{k-1})_j| \leq eps \min\{|(\mathbf{y}_k)_j|, |(\mathbf{y}_{k-1})_j|\}$$

for all components j, then clearly the two iterates do not differ with respect to machine precision, hence the method stagnates, that is $\mathbf{y}_k = \mathbf{y}_{k+1} = \ldots$. This effect can appear when the requirements are too strong and the difference between the exact solution and the approximation is small, so convergence is achieved. However, this effect can also appear considerably earlier due to slow convergence.

3.7.2 Stopping conditions

Obviously the main stopping condition for the inner iteration is due to successfully reaching the specified convergence level for the inner iterations. Further it is sensible to stop the inner iterations, when the convergence level for the outer method, here the eigenvalue residual, is reached. However, the inner method might fail to provide a satisfactory answer and create additional cost by trying repeatedly without success. Hence, one likes to detect any failure as quickly as possible and return to the outer method, hopefully still providing a sensible answer and a flag indicating the kind of failure.

The MatLab routine MINRES checks in each iteration for stagnation, that is if the round off level is reached. However our experience is that this condition is too tight and creates additional costs due to late detection of failure. Additionally due to the late detection the solution is often not a good approximation to the sought eigenvalue.

The stopping condition from Simoncini and Eldén (2002) introduced earlier, (3.47) can be used either instead of the residual condition, so as main stopping condition or as an additional condition to gain more reliability. However our practical experience is not in favour of either of these variations. If one knows the correct stopping parameter, τ_{SE} in (3.47), both variations work excellently. However, in our experience this parameter τ_{SE} depends on the matrix, the considered eigenvalue and on the starting vector $\mathbf{x}^0$ and so far no estimator for a good choice of τ_{SE} is available. For more on this see Example 3.5 in Section 3.8.3.

More satisfactory especially in case of PInvit, is to check the outer convergence condition in each iteration of MINRES. This has the advantage that specially for almost singular systems convergence can be detected before MINRES derails. However this is not the case if the convergence condition is too tight (too strong requirements). For this case an additional stopping condition is needed. As explained earlier in this case the residual norm estimator and the residual norm behave differently.

In the following we describe how the residual stopping condition can be implemented without needing further matrix vector products or application of the preconditioner. The same technique can be used to check the eigenvalue residual norm. Then we discuss how the additional stopping condition can be implemented.

Residual Stopping Condition

In Section 3.6.5 we discussed a few aspects of implementing MINRES. Here we repeat some of these ideas in order to understand how these stopping conditions can be calculated without needing further matrix vector products or applications of the preconditioner.

To make this discussion applicable for all methods discussed so far denote the preconditioner by $P = P_1 P_2$ where P_1 or P_2 might be the identity matrix.

MINRES calculates a basis U for the Krylov subspace using the Arnoldi / Lanczos sequence. An essential ingredient is a three term recurrence formula to construct U and T such that $P_1^{-1}(A-\sigma I)P_2^{-1}U = UT$. Further $T = QR$ where R upper triangular with only three diagonal not equal to zero. Another important ingredient of the MINRES algorithm is a three term recurrence formula for R^{-1} which can be applied to any matrix B to obtain BR^{-1}. As in Section 3.6.5 where we explained how $P\mathbf{y}$ can be calculated, the calculation for any expression of the form $\mathbf{z} = B\mathbf{y} = (BR^{-1})Q^T\mathbf{e}_1$ is likewise as long as B is available. To calculate $(A-\sigma I)\mathbf{y}$ we need only four vectors and the recurrence formula for R^{-1}. Hence computing the residual $\|\mathbf{b}^i - (A-\sigma^i I)\mathbf{y}\|$ involves only a few vector additions and one scalar product. In case of the approach by Simoncini and Eldén (2002) we have the residual $\|P_1^{-T}\mathbf{b}^i - P_1^{-T}P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{y}^i\|$. To calculate $P_1^{-T}P_1^{-1}(A-\sigma^i I)P_1^{-T}\mathbf{y}^i$ without using any further matrix vector products or application of the preconditioner we need to reorder a few steps in the algorithm and storage for a few additional vectors.

Eigenvalue Residual

As explained above the calculation of $(A-\sigma I)\mathbf{y}$ does not need to cost any matrix vector products. Using $(A-\sigma I)\mathbf{y}$ we can calculate the eigenvalue residual by

$$\frac{\|A\mathbf{y} - \varrho(\mathbf{y})\mathbf{y}\|}{\|\mathbf{y}\|} = \frac{\|(A-\sigma I)\mathbf{y} - \dfrac{\mathbf{y}^T(A-\sigma I)\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\mathbf{y}\|}{\|\mathbf{y}\|}.$$

However it is appropriate to check the eigenvalue residual, or the scaled eigenvalue residual as convergence conditions for the outer iteration. If the smallest eigenvalue of A is larger then a better approximation of $|s^i|$ is given by the scaled eigenvalue residual

$$s^i_{est} := \frac{\|A\mathbf{y} - \dfrac{\mathbf{y}^T A\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\mathbf{y}\|}{\|\mathbf{y}\||\sigma|}.$$

Smit and Paardekooper (1999) shows that $|s^i||c^i| \leq s^i_{est}$ for σ equaling the smallest eigenvalue. The eigenvalue residual is appropriate if the sought eigenvalue has an absolute value of order 1 or smaller while $\|A\| > 1$.

Additional Stopping Condition

Preconditioned MINRES calculates an estimator for $\|\mathbf{res}^i_k\|_{P^{-1}}$ called *snprod* as part of updating the solution $\mathbf{y}^i_k$. When round off errors lead to a loss of convergence this estimator still converges towards zero. Based on this fact we could stop the inner iterations when *snprod*/ $\|\mathbf{res}^i_k\|_{P^{-1}}$ drops below a threshold. However this would require the calculation of $\|\mathbf{res}^i_k\|_{P^{-1}}$. Further the decrease of *snprod*/ $\|\mathbf{res}^i_k\|_{P^{-1}}$ can

also be caused by failing to calculate $\|\mathbf{res}_k^i\|_{P^{-1}}$ accurately enough. Therefore we look for a different stopping condition. Assuming *snprod* reflects $\|\mathbf{res}_k^i\|$ better than the calculated one we can use the fact

$$\|\mathbf{res}_k^i\|_{P^{-1}} \leq \|P^{-1}\| \|\mathbf{res}_k^i\|_2 .$$

We are interested in stopping when $\|\mathbf{res}_k^i\| \leq \tau^i$, therefore we might use $snprod \leq const\,\tau$ as additional stopping condition. In our tests we used $snprod \leq 10^{-5}\tau \, \|\mathbf{res}_0^i\|_{P^{-1}}$ with good results. However the constant 10^{-5} is neither optimal nor independent of the eigenvalue problem, though changes in this constant had little effect. The condition also works well when *snprod* does not reflect $\|\mathbf{res}_k^i\|_{P^{-1}}$.

3.8 Numerical Examples

In this section we illustrate the theoretical results on the efficiency for inexact inverse iteration using unpreconditioned MINRES and preconditioned MINRES. Further we illustrate the convergence and efficiency of the approach Rüde and Schmid (1995) as well as of methods using the modified equation (3.46). We report on tests where we compared the performance of various methods. As a benchmark to compare our algorithm against we use LOBPCG, see Knyazev (2000). Also tests against linear solves using MINRES are done to illustrate what the cost factor between solving an eigenvalue problem and a linear solve is. To illustrate the convergence we consider as in Chapter 2 the 'Poisson' eigenvalue problem and the matrix 'bcsstk09' form Matrix-Market (`http://gams.cam.nist.gov/MatrixMarket/index.html`).

The examples and some useful abbreviations are introduced in Section 3.8.1. In Section 3.8.2 we discuss the results on the numerical test with respect to the efficiency for the methods considered in Sections 3.3 and 3.4, that are Invit, RQIf and RQId. Then in Section 3.8.3 we discuss the performances of the methods analysed in Sections 3.5 and 3.6, that are SE, all variations of the inverse correction method and PInvit. The discussion in Section 3.8.3 will be with respect to convergence and efficiency. Finally in Section 3.8.4 we summarise our theoretical findings and our practical experience.

3.8.1 Notation and examples

We introduced some abbreviations in Chapter 2. We now extend this list to all methods compared here.

Invit stands for inverse iteration with fixed shift and decreasing tolerance, $\sigma^i = \varrho^0$ and $\tau^i = \min\{\widetilde{C}_2 \, |\varrho^i|^{-1}\|r^i\|_2, \tau_0\}$.

RQIf is the Rayleigh quotient iteration with fixed tolerance, $\sigma^i = \varrho^i$ and $\tau^i = \tau_0$.

RQId is the Rayleigh quotient iteration with decreasing tolerance, $\sigma^i = \varrho^i$ and $\tau^i = \min\{\widetilde{C}_2\ |\varrho^i|^{-1}\|r^i\|_2, \tau_0\}$.

SE stands for the approach by Simoncini and Eldén (2002) using the alternative system (3.46), with $\sigma^i = \varrho^i$ and the stopping condition $\|\mathbf{y}_k\| \geq \|\mathbf{r}^i\|^{-1}$ and simultaneously $(\|\mathbf{y}_k\| - \|\mathbf{y}_{k-1}\|)\ \|\mathbf{y}_k\|^{-1} \leq \tau_{SE}$.

PInvit stands for our variation of the approach of Simoncini and Eldén (2002), that is Inverse Iteration using the alternative system (3.46) with a residual stopping condition, $\sigma^i = \varrho^i$ and fixed tolerance $\tau^i = \tau_0$. Here we use the fact that a Cholesky preconditioner is used in MINRES and use P to calculate $\mathbf{b} = P\mathbf{x}$.

PInvit+ that is Algorithm 4 with $\sigma^i = \varrho^i$ and $\tau^i = \tau_0$. Here we ignore the fact that a Cholesky preconditioner is used. The first iteration is a standard step of inexact inverse iteration like Invit, RQIf and RQId, but calculating additionally $P\mathbf{y}$ to provide $\mathbf{b} = P\mathbf{x}$. The remaining iterations are as in PInvit.

ICMf Inverse Correction with fixed shift, first iteration is inexact inverse iteration, Algorithm 2 with $\sigma^0 = \varrho^0$, then inverse correction, Algorithm 3 with $\sigma^i = \varrho^0$ and $\tau^i = \tau_0$.

ICMfp Inverse Correction with fixed perturbation, first iteration is inexact inverse iteration, Algorithm 2 with $\sigma^0 = \varrho^0$, then inverse correction, Algorithm 3 with $\sigma^i = \varrho^i + C_9$ and $\tau^i = \tau_0$.

ICMlp Inverse Correction with linear perturbation, first iteration is inexact inverse iteration, Algorithm 2 with $\sigma^0 = \varrho^0$, then inverse correction, Algorithm 3 with $\sigma^i = \varrho^i + C_9\mathbf{r}^i / |\varrho^i|$ and $\tau^i = \min\{\frac{1}{2}, \widetilde{C}_2\ |\varrho^i - \sigma^i| \ / \ |\varrho|\}$.

ICMqp Inverse Correction with quadratic perturbation, first iteration is inexact inverse iteration, Algorithm 2 with $\sigma^0 = \varrho^0$, then inverse correction, Algorithm 3 with $\sigma^i = \varrho^i + C_9\ \|\mathbf{r}^i\|^2\ /(\varrho^i)^2$ and $\tau^i = \min\{\frac{1}{2}, \widetilde{C}_2\ |\varrho^i - \sigma^i| \ / \ |\varrho|\}$.

LOBPCG is an algorithm by Knyazev (2001), the code can be downloaded from `http://www-math.cudenver.edu/~aknyazev/software/CG`.

Poisson Poisson eigenvalue problem on a rectangular domain, aspect ratio 1/1.3, with Dirichlet boundary conditions. For discretisation we use thirteen grid points per direction and a second order central finite difference scheme. We consider only the smallest eigenvalue of this 121 × 121 matrix,

i^{th} smallest	1	2	121
value	15.6	32.6	901.2

bcsstk09 We use the real symmetric matrix bcsstk09 from Matrix-Market. All tests presented have the same starting vector $\mathbf{x}^0 = c^0\mathbf{v}_1 + s^0\mathbf{w}^0$ where $t^0 = s^0/c^0 = 0.02$. We consider two different strategies. One where we try to find the smallest eigenvalue and the other where we try to find the 20^{th} smallest eigenvalue of the 1083×1083 matrix. In the following table we summarise those eigenvalues which describe the difficulty for the corresponding tests.

i^{th} smallest	1	2	19	20	21	1083
value	7.1e+3	2.7e+4	3.7e+5	4.1e+5	4.4e+5	6.7e+7

3.8.2 Standard approaches

Here we illustrate the efficiency results for the methods analysed in Sections 3.3 and 3.4, that are Invit, RQIf and RQId. Besides the here presented tests we did tests with different starting vectors, different parameter settings, different stopping conditions, and different eigenvalue problems. As we used in the theory the tangents as a measure for convergence we do so here, knowing that the tangents is in practise not available. Here we use 12-15 digit accurate approximations of the sought eigenvector to calculate the tangent, hence we use 10^{-10} as targeted accuracy level, so that the comparison solution should be more accurate than the current iterate $\mathbf{x}^i$.

To illustrate Theorems 3.10 and 3.13 we consider three test on the example 'bcsstk09'. The first will use the three practical methods while the second one uses arbitrary shifts σ^i and tolerance conditions τ^i in order to demonstrate the quality of the bounds for the number of inner iterations per outer iteration $\mathcal{L}^i$. The third test adapts the idea of the second test to the total number of inner iterations $\mathcal{T}$. Finally to support Lemma 3.11 we use the example 'Poisson' and variable precision arithmetic to illustrate the difference between the quadratic convergence of RQIf and the cubic convergence of RQId.

Test 3.1 *We use Invit, RQIf and RQId on 'bcsstk09' together with unpreconditioned MINRES as well as with preconditioned MINRES. As a preconditioner we use an incomplete Cholesky factorisation of the matrix A. To calculate the preconditioner we use the MatLab routine* cholinc *with droptol* $= 0.01$. *In all test runs we try to find the 20th smallest eigenvalue to an accuracy of* $t^N \approx 10^{-10}$. *As a stopping condition for the outer method we use* $\|\mathbf{r}^i\|_2 \;/\; |\varrho^i| \leq 10^{-10}$. *For the inner iterations we use the stopping conditions as discussed in Section 3.7.2. The parameters used are given together with the convergence and efficiency data in Tables 3.1, 3.2, 3.3 and 3.4.*

Test 3.1 repeats the Tests 2.1, 2.2 and 2.3, but with different starting vectors. Here we table the number of inner iterations $\mathcal{L}^i$ instead of the convergence progress t^i/t^{i+1}. The main parameter $\widetilde{C}_2$ is for both tests in Tables 3.1 the same. However the smaller value for τ_0 enforces that $\|\mathbf{r}^i\||s^i|^{-1}$ is marginally smaller. The resulting increase in the

	$\tau^0 = 0.2\ \widetilde{C}_2 = 100$			$\tau^0 = 0.02\ \widetilde{C}_2 = 100$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.5e+05	1.0e-02	35	2.5e+05	1.0e-02	154
1	3.1e+02	1.3e-03	37	3.5e+01	6.3e-04	168
2	1.2e+02	9.8e-04	64	1.3e+01	2.8e-04	170
3	4.6e+01	7.3e-04	157	5.1e+00	9.7e-05	181
4	1.8e+01	4.3e-04	150	1.9e+00	3.9e-05	182
5	6.9e+00	1.1e-04	152	7.5e-01	1.5e-05	183
6	2.7e+00	6.5e-05	148	2.8e-01	5.5e-06	184
7	1.0e+00	1.7e-05	142	1.1e-01	2.2e-06	185
8	4.1e-01	1.0e-05	140	4.0e-02	7.5e-07	180
9	1.6e-01	2.7e-06	119	1.6e-02	3.4e-07	181
10	6.2e-02	1.6e-06	122	6.0e-03	1.1e-07	165
11	2.4e-02	3.8e-07	98	2.3e-03	5.3e-08	169
12	9.5e-03	2.4e-07	134	8.9e-04	1.6e-08	156
13	3.5e-03	4.6e-08	97	3.5e-04	8.4e-09	163
14	1.4e-03	3.4e-08	142	1.3e-04	2.3e-09	140
15	4.9e-04	6.3e-09	117	5.3e-05	1.3e-09	83
16	1.9e-04	4.8e-09	135	4.0e-05	9.8e-10	
17	7.4e-05	1.0e-09	61			
18	4.1e-05	9.2e-10				
$\mathcal{T}$			2050			2644

Table 3.1: Invit using MINRES on 'bcsstk09' (Test 3.1)

	$\tau^0 = 0.05$			$\tau^0 = 0.2\ \widetilde{C}_2 = 2$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	5.0e+05	2.0e-02	55	5.0e+05	2.0e-02	27
1	3.4e+02	2.1e-03	408	1.2e+03	3.7e-03	408
2	3.2e-04	4.9e-10	27	3.6e-03	2.3e-08	268
3	4.0e-05	3.0e-10		3.8e-05	5.6e-10	
$\mathcal{T}$			490			703

Table 3.2: RQIf and RQId using MINRES on 'bcsstk09' (Test 3.1)

	$\tau^0 = 0.2\ \widetilde{C}_2 = 100$			$\tau^0 = 0.02\ \widetilde{C}_2 = 100$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.5e+05	1.0e-02	45	2.5e+05	1.0e-02	51
1	1.6e+02	2.6e-03	46	2.9e+01	1.2e-04	54
2	4.6e+01	5.6e-04	53	6.8e+00	2.5e-05	59
3	1.2e+01	5.9e-05	54	1.1e+00	2.4e-06	61
4	4.3e+00	1.7e-05	59	4.2e-01	8.2e-07	65
5	1.2e+00	2.8e-06	62	5.2e-02	1.6e-07	69
6	4.2e-01	8.5e-07	64	1.2e-02	1.4e-08	73
7	1.6e-01	3.4e-07	67	9.9e-04	2.5e-09	77
8	2.1e-02	6.9e-08	71	3.4e-04	5.3e-10	78
9	4.4e-03	5.0e-09	73	8.4e-05	2.8e-10	80
10	1.1e-03	2.1e-09	74	3.7e-05	1.2e-10	
11	3.9e-04	1.4e-09	78			
12	6.3e-05	2.1e-10	80			
13	3.0e-05	1.2e-10				
$\mathcal{T}$			826			667

Table 3.3: Invit using prec. MINRES on 'bcsstk09' (Test 3.1)

	$\tau^0 = 0.1$			$\tau^0 = 0.2\ \widetilde{C}_2 = 2$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	5.0e+05	2.0e-02	35	5.0e+05	2.0e-02	32
1	6.2e+02	1.3e-02	65	1.0e+03	1.6e-02	66
2	6.2e-02	4.2e-07	86	1.2e-01	3.1e-06	85
3	2.3e-05	9.7e-11		9.3e-06	3.1e-11	
$\mathcal{T}$			186			183

Table 3.4: RQIf and RQId using prec. MINRES on 'bcsstk09' (Test 3.1)

number of inner iterations $\mathcal{L}^i$ is predicted by the a posteriori bound (3.20). Further we observe that the unpreconditioned MINRES solves are too expensive to make this linear converging method worthwhile. From Tables 3.1 and 3.2 we see that the number of inner iterations is related to the reduction of the tangents. Further by comparing Tables 3.1 and 3.2 we observe that the number of inner iterations per outer iteration increases for the variable shift techniques RQIf and RQId. This is expected as the a priori bounds (3.24) and (3.32) are linearly increasing in $\log(|\lambda_1 - \sigma^i|^{-1})$. However this increase should not concern as the outer convergence accelerates and by that the total number of inner iterations decreases such that the total number of inner iterations $\mathcal{T}$ is reduced.

Comparing RQIf and RQId we observe that the convergence of the two approaches is almost indistinguishable. More important is that both algorithm need the same number of outer iterations. For this case the a posteriori bound (3.22) on the total number of outer iterations $\mathcal{T}$ differs only in the constants C_3 and C_8. The chosen value for C_2 in RQIf enforces that C_8 is larger than for RQId. While in the second iteration for RQIf and RQId, see Table 3.2, $\mathcal{L}^i$ is the same, the progress t^{i+1}/t^i differs. Therefore the observed better performance of RQIf over RQId is supported by our bound on $\mathcal{T}$, (3.22). Before we draw more attention to the differences between RQIf and RQId we consider a test illustrating the quality of the bounds for $\mathcal{L}^i$ and $\mathcal{T}$. This test will also give a better inside to the role of C_8.

Test 3.2 *Consider the example 'bcsstk09'. We use inexact inverse iteration with different parameter choices, satisfying the conditions*

$$|s^i| \leq C_8 \, \|\mathbf{res}^i_{k^i}\| \quad \textit{and} \quad \|\mathbf{res}^i_{k^i}\| \leq \tau^i, \tag{3.64}$$

where τ^i such that the conditions of Theorems 3.10 and 3.13 are satisfied. As a linear solver we use unpreconditioned and preconditioned MINRES. We consider the 20th smallest eigenvalue and perform always only one outer iteration. We restart this test with different starting vectors, with different approximation accuracies t^0 and with different error directions $\mathbf{u}^0$. Then for the unpreconditioned case, Figure 3-1, we plot the number of inner iterations $\mathcal{L}^i$ performed in this one outer iteration against the progress that was achieved t^i/t^{i+1}. The preconditioned case is illustrated in Figure 3-2 where we plot the number of inner iterations against the achieved approximation quality $1/t^{i+1}$.

In Figure 3-1 and 3-2 we used black asterisk for successful test runs with $C_8 \geq 1$. For successful runs with $0.01 \leq C_8 < 1$ we used green dots, the red dots represent tests were MINRES suffered a breakdown, which occurred always with $t^{i+1} \approx 10^{-12}$. In Figure 3-1 we give additional to the results marked by the asterisks and dots, two lines representing the slope of the bound for two different sets of Γ. The lower one corresponds to a set Γ with $|\Gamma| = 100$ while for the upper one $|\Gamma| = 200$. The constants

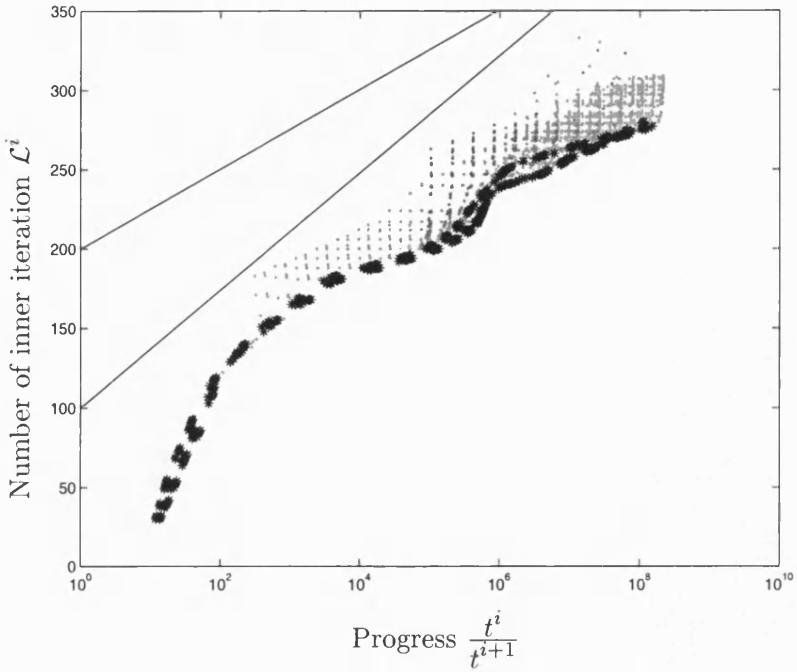


Figure 3-1: arbitrary shifts and tolerance constraints, unpreconditioned MINRES (Test 3.2)

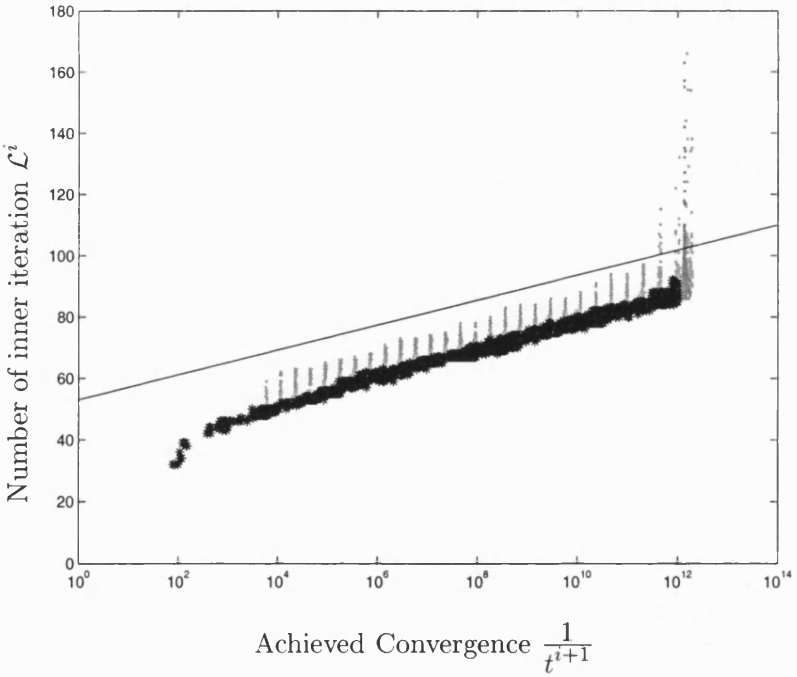


Figure 3-2: arbitrary shifts and tolerance constraints, preconditioned MINRES (Test 3.2)

p_Γ are in this case so large that the we only indicated the slope of the bound, while the actual bound is meaningless. More interesting is that by allowing $1 \leq C_8 \leq 100$, see green dots in Figure 3-1, the empirical slope moves upwards. This behaviour is as such predicted by the a posteriori bound on $\mathcal{L}^i$, (3.20), which states $\mathcal{L}^i \propto \log(1 + C_8)$. As for this test no additional stopping conditions were used we observed breakdowns of the linear solver, those of them lying inside the graph are plotted in red. However there are more such failures outside of the graph.

So far we made no comments on the preconditioned case nor on the differences between the preconditioned and the unpreconditioned case.

In Test 3.1 we used an incomplete Cholesky factorisation with droptol = 0.01. This choice of the drop-tolerance leads to the cost relation

$$\frac{\text{cost of applying the preconditioner}}{\text{cost of applying A}} = 1.64.$$

Introducing such a preconditioner forces the cost per iteration to rise by not more than a factor three. So to gain a similar performance for the preconditioned approach the number of inner iterations should be a third. While the number of inner iterations reduces by about one third for Invit and therefore not an improvement, the reduction is better for RQIf and RQId. Further we observe that in the preconditioned case the cost of the linear solve is related to $1/t^{i+1}$, see for example Figure 3-2. Again the asterisks denote test runs with $C_8 \leq 1$ and the green dots with $1 \leq C_8 \leq 100$. Another difference is the quality of the bound on $\mathcal{L}^i$. The line in Figure 3-2 represents the actual bound for $\mathcal{L}^i$, (3.28), using p_Γ and q_Γ according to Lemma 3.5 where $D = D_\Gamma$ as defined in (3.19) with $\Gamma = \{1, 2, \ldots, 20\}$, where $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{20} \leq \lambda_j$ for all $j > 20$.

We now repeat Test 3.2 but calculate always until $t^N \leq t^0 10^{-\gamma}$ is reached.

Test 3.3 *Consider the example 'bcsstk09' and compute its 20th smallest eigenvalue. We use inexact inverse iteration with different parameter choices, satisfying the conditions* $| s^i | \leq C_8 \, \| \mathbf{res}^i \|$ *and* $\| \mathbf{res}^i \| \leq \tau^i$, *where* τ^i *is such that the conditions of Theorem 3.13 are satisfied. As a linear solver we use unpreconditioned and preconditioned MINRES. The preconditioner is constructed using the MatLab routine* cholinc *with droptol* $= 0.01$. *We restart this test with different initial error direction* $\mathbf{u}^0$ *while* $t^0 = 0.01$ *is fixed. In Figures 3-3, unpreconditioned MINRES, and 3-4, preconditioned MINRES, we plot the total number of inner iterations* $\mathcal{T}$ *against the number of outer iterations* $\mathcal{N}$ *for each run.*

In Test 3.3 each test run choses σ^i and τ^i randomly. The chosen values are then check against the convergence conditions of Theorem 3.13 and if necessary rejected. After running inexact inverse iteration using one set of parameters a single entry in the graph is made. The colour of the entry depends on the value of C_8, for $C_8 <$ 0.01 black, for $0.01 \leq C_8 < 0.1$ magenta, for $0.1 \leq C_8 < 1$ red and all others blue

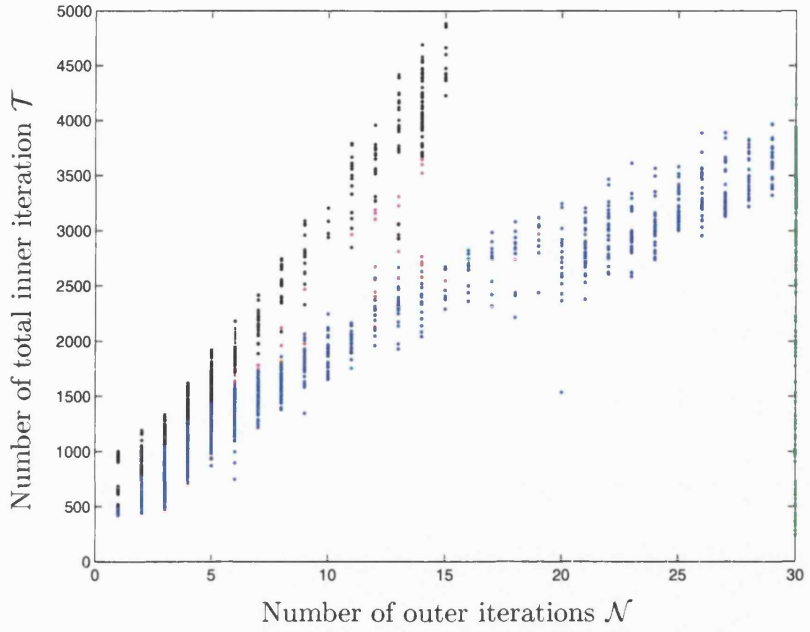


Figure 3-3: arbitrary shifts and tolerance constraints, preconditioned MINRES (Test 3.3)

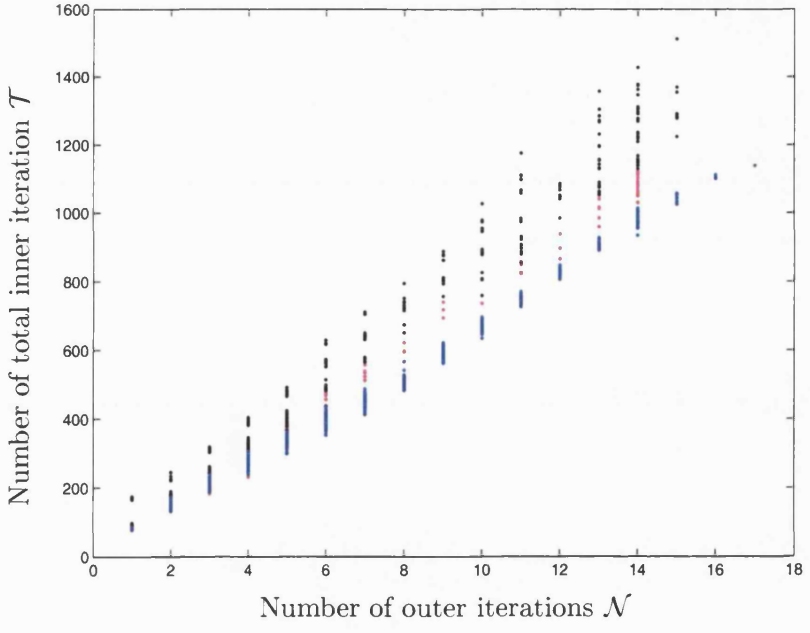


Figure 3-4: arbitrary shifts and tolerance constraints, preconditioned MINRES (Test 3.3)

	RQIf		RQId	
i	$\log_{10} t^i$	k^i	$\log_{10} t^i$	k^i
0	-0.12	19	-0.14	15
1	-1.41	19	-1.62	24
2	-3.85	33	-4.33	45
3	-9.03	50	-12.90	78
4	-19.46	76	-36.19	113
5	-40.72	108	-82.66	
6	-82.96			
		305		275

Table 3.5: Cubic is better than quadratic (Test 3.4)

unless convergence was not reached in the first 30 outer iterations (green dots). The resulting graph for the unpreconditioned case, Figure 3-3, illustrates that the bound on $\mathcal{T}$ depends linearly on the number of outer iterations. In contrast in the preconditioned case the total number of inner iterations $\mathcal{T}$ grows faster than linear in the number of outer iterations, despite the appearance in Figure 3-4. As Figure 3-3 indicates using inexact linear solves one can obtain 'lucky' performances with small total cost $\mathcal{T}$ while a large number of outer iterations was performed. However to ensure low costs one has to cut down the number of outer iterations $\mathcal{N}$.

Earlier we diverted our attention from the difference in the performance of RQIf and RQId. As both algorithm need only three outer iterations in test (3.1) the discrete nature of $\mathcal{N}$ is not negligible. Therefore γ is too small to apply Lemma 3.11. In contrast with Test 2.4 using variable precision arithmetic (vpa) we had a test where γ was large enough. Here we repeat this test and discuss it with respect to the efficiency.

Test 3.4 *We use RQIf and RQId on 'Poisson' together with unpreconditioned MINRES. In all test runs we try to find the smallest eigenvalue to an accuracy of* $t^N \approx 10^{-80}$. *We stop the inner iterations due to either reaching the required residual tolerance or the required tangent* t^N. *The results are presented in Table 3.5.*

As Table 3.5 shows $\mathcal{N}$ differs only by one between RQIf and RQId, but this is already enough to obtain a lower number of total inner iterations $\mathcal{N}$ for RQId than for RQIf. However as this example illustrates γ needs to be large in order to ensure a reduction in $\mathcal{N}$ for RQId over RQIf. The difference between the two methods becomes more apparent when a preconditioner is used. Nevertheless even for the preconditioned case γ needs to be large in order to neglect the discrete nature of $\mathcal{N}$. So for practical situations with limited machine precision we can expect RQIf and RQId to perform similarly.

	extreme $\tau_{SE} = 0.1$			interior $\tau_{SE} = 10^{-4}$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	4	5.0e+05	2.0e-02	12
1	1.2e+04	4.6e-03	10	1.1e+04	7.0e-03	57
2	6.9e+01	5.1e-05	18	4.9e+00	2.1e-05	60
3	3.1e-03	2.0e-09	13	3.6e-05	9.6e-11	
4	4.8e-07	2.7e-12				
$\mathcal{T}$			45			129

Table 3.6: SE using MINRES on 'bcsstk09' (Test 3.5)

	extreme $\tau^0 = 0.5$			interior $\tau^0 = 0.5$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	2	5.0e+05	2.0e-02	4
1	1.5e+04	6.5e-03	7	2.2e+04	7.7e-03	38
2	1.7e+02	3.6e-04	15	4.7e+01	2.6e-04	60
3	3.9e-02	4.5e-08	17	1.5e-03	5.9e-09	32
4	4.1e-07	4.0e-12		3.4e-05	4.0e-10	
$\mathcal{T}$			41			134

Table 3.7: PInvit using MINRES on 'bcsstk09' (Test 3.5)

3.8.3 Variations of Inexact Inverse Iteration

In the following we compare the variations of inexact inverse iteration as introduced in Section 3.5 and 3.6. We will compare them with RQIf and RQId as illustrated in the previous section. Further LOBPCG and a simple inexact linear solve using MINRES are considered as benchmarks later on. To allow a fair comparison between the methods we use the same initial approximation and the same preconditioner for all methods. However we made many more tests with other starting vectors, other parameter values, and other matrices.

Test 3.5 *Consider example 'bcsstk09' and apply SE, PInvit and PInvit+ to the smallest and the 20th eigenvalue of 'bcsstk09' using unpreconditioned MINRES as linear solver.*

We recall that SE and PInvit use the preconditioned alternative system (3.46) either explicitly or implicitly, hence a single inner iteration of unpreconditioned MINRES has the same cost as a single inner iteration of preconditioned MINRES on $(A - \sigma I)\mathbf{y} = \mathbf{b}$. Therefore the methods to compare are RQIF and RQId using preconditioned MINRES, see Table 3.4. Comparing the results of RQIf and RQId with those for SE, Table 3.6 we observe the superiority of the approach from Simoncini and Eldén (2002). For all methods we compare here except LOBPCG, SE provided the optimal results. However these optimal performances were dependent on the eigenvalue, the chosen value for τ_{SE}

	extreme eigenvalue $\alpha = 0.05$			interior eigenvalue $\alpha = 0.05$		
	$\| r \|$	t^i	$\mathcal{L}^i$	$\| r \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	13	5.0e+05	2.0e-02	45
1	4.5e+01	3.1e-04	19	1.6e+02	2.0e-03	55
2	1.8e-02	9.3e-09	16	1.1e-01	4.8e-07	50
3	5.5e-07	1.93e-12		2.0e-05	6.47e-11	
			48			150

Table 3.8: PInvit+ using prec. MINRES on 'bcsstk09' (Test 3.5)

and the initial vector $\mathbf{x}^0$. Optimal results for the extreme eigenvalue were obtained with $\tau_{SE} \approx 0.2$, the actual choice depended on the initial vector. In contrast for the interior eigenvalue choices of $\tau_{SE} > 0.01$ lead to divergence. According to our experience this effect is independent of the quality of the initial approximation $\mathbf{x}^0$. To obtain excellent performances for the interior eigenvalue we used $\tau_{SE} = 0.0005$. It is not a cure to tighten the condition τ_{SE} a priori as than the cost increases dramatically specially for the interior eigenvalue problem.

SE and PInvit differ only in the stopping condition for the inner iterations. Therefore the also excellent results for PInvit highlights the benefit of considering the alternative update equation (3.46). In contrast to SE, PInvit is very robust with respect to the stopping condition. Choosing $\tau^i = 0.1$ instead of $\tau^i = 0.5$ as in Table 3.7 increases the overall cost $\mathcal{T}$ by 5 iterations for the extreme eigenvalue and by 30 iterations for the interior eigenvalue. So again the optimal performance is sensitive to the choice of the stopping condition. However for PInvit this optimal choice is independent of the initial approximation and independent of the eigenvalue.

PInvit+ using MINRES with (incomplete) Cholesky preconditioning is from the second iteration onwards the same as PInvit. As the first iteration of PInvit+ is the same as for Invit, RQIF and RQId, using the standard right hand side, we expect PInvit+ to be inferior to PInvit. However we expect PInvit+ to be sufficiently cheaper than RQIf and RQId.

Another advantage of SE, PInvit and PInvit+ is the robustness these three methods have over RQIf and RQId when MINRES is used with the additional stopping conditions discussed in Section 3.7. Invit, RQIf and RQId suffer from breakdowns for tight outer convergence conditions $\|\mathbf{r}^i\| \,/\, |\varrho^i| \leq 10^{-10}$. This is not the case for SE and PInvitunless the requirements are too strong. Further when the stopping condition is too tight to be ever satisfied, for example $\|\mathbf{r}^i\| \,/\, |\varrho^i| \leq 10^{-14}$ this is detected in SE and PInvit and a good approximation with $\|\mathbf{r}^i\| \,/\, |\varrho^i| \leq 10^{-12}$ can be provided. In contrast Invit, RQIf and RQId using preconditioned MINRES break down without providing a highly accurate approximation to the sought eigenpair.

	extreme $\tau^0 = 0.5$			interior $\tau^0 = 3.9 * 10^{-3}$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	13	2.5e+05	1.0e-02	49
1	4.5e+01	3.1e-04	4	5.2e+01	1.8e-04	45
2	9.4e+00	7.0e-05	4	3.3e-01	4.0e-06	47
3	2.0e+00	1.6e-05	5	6.4e-03	1.4e-07	48
4	3.6e-01	9.9e-07	3	2.2e-04	5.7e-09	47
5	4.7e-02	3.8e-07	4	9.0e-06	2.3e-10	
6	9.8e-03	3.4e-08	3			
7	1.8e-03	1.5e-08	4			
8	3.4e-04	1.1e-09	3			
9	6.6e-05	4.2e-10	4			
10	1.1e-05	8.3e-11	5			
11	2.2e-06	9.9e-12	3			
12	4.2e-07	3.9e-12				
$\mathcal{T}$			55			236

Table 3.9: ICMf using prec. MINRES on 'bcsstk09' (Test 3.6)

	extreme $\tau_0 = 1.4 * 10^{-3}$ $pert = 1$			interior $\tau_0 = 2.5 * 10^{-5}$ $pert = 1$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	13	5.0e+05	2.0e-02	45
1	4.5e+01	3.1e-04	13	2.2e+02	2.6e-03	84
2	4.7e-02	1.4e-07	12	2.9e-03	5.8e-08	57
3	6.2e-05	4.8e-10	13	6.9e-08	1.6e-12	
4	4.7e-08	7.9e-14				
$\mathcal{T}$			51			186

Table 3.10: ICMfp using prec. MINRES on 'bcsstk09' (Test 3.6)

Test 3.6 *We repeat Test 3.5 with the methods ICMf, ICMfp,ICMlp, and ICMqp. The results are presented in Tables 3.9-3.12.*

In general the cost per outer iteration reduces for the inverse correction method in comparison to Invit, RQIf and RQId and also in comparison to SE and PInvit. However ICMf needs too many outer iterations so that the advantage of $\mathcal{L}^i$ being small for all i does not pay off. In case of the interior eigenvalue this is even more apparent. However ICMf is the most robust of all the here tested methods.

A major concern for ICMfp, ICMlp and ICMqp is the erratic convergence behaviour. This erratic convergence behaviour leads to a lack of control in the outer iteration. As a result ICMfp, ICMlp and ICMqp suffer frequently from breakdowns of MINRES. While well in advance of a breakdown the eigenvalue residual relates to the tangent as $\|\mathbf{r}^i\| \, (t^i)^{-1} \approx 10^5$, in the breakdown situation the relation becomes $\|\mathbf{r}^i\| \, (t^i)^{-1} \approx 10^7$. This makes it even harder to detect a breakdown. However the occasionally good

	extreme $\widetilde{C}_2 = 10$ $C_9 = 1$			interior $\widetilde{C}_2 = 1$ $C_9 = 0.01$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	13	5.0e+05	2.0e-02	45
1	4.5e+01	3.1e-04	31	2.2e+02	2.6e-03	184
2	2.4e-04	2.7e-10	87	7.9e-04	1.2e-08	
3	1.4e-04	3.4e-12				
$\mathcal{T}$			131			229

Table 3.11: ICMlp using prec. MINRES on 'bcsstk09' (Test 3.6)

	extreme $\widetilde{C}_2 = 10$ $C_9 = 1$			interior $\widetilde{C}_2 = 1$ $C_9 = 0.01$		
i	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$	$\| \mathbf{r}^i \|$	t^i	$\mathcal{L}^i$
0	2.7e+05	1.0e-02	13	5.0e+05	2.0e-02	45
1	4.5e+01	3.1e-04	65	2.2e+02	2.6e-03	188
2	1.6e-04	1.6e-10		1.1e-0r2	1.1e-08	
$\mathcal{T}$			78			233

Table 3.12: ICMqp using prec. MINRES on 'bcsstk09' (Test 3.6)

performance gets close to the quality of SE and PInvit.

To demonstrate the quality of the performance of RQIf, RQId, SE, PInvit and PInvit+ we now introduce two benchmarks.

Test 3.7 *Consider example 'bcsstk09' and apply LOBPCG using preconditioned MINRES to the smallest eigenvalue. We repeat this for different right-hand sides and two targeted eigenvalue accuracies,* 10^{-10} *and* 10^{-12}*. Further we apply these right-hand sides also to RQIf, RQId, PInvit and PInvit+. The results are given in Table 3.13.*

LOBPCG (Locally Optimal Block Preconditioned Conjugate Gradient method) is a one-level-method to find the smallest eigenvalue of a symmetric positive definite matrix A. This algorithm is one of the most efficient algorithms for this task, see Knyazev (2000). For more detail on LOBPCG see also Knyazev and Neymeyr (2003). Here we use LOBPCG with blocksize $k = 1$ and a 3 dimensional subspace to calculate the eigenpair approximation by a Rayleigh-Ritz analysis. This should lead to a considerably better approximation than the RQ of the current iterate. Therefore we expect that

	10^{-10}	10^{-12}
RQIf	68	70
RQId	63-69	66-73
PInvit	42	47-52
PInvit+	47	52-60
LOBPCG	34-37	38-41*

Table 3.13: Comparison of over all cost $\mathcal{T}$, Test 3.7

relative accuracy	unprec. MINRES	prec. MINRES, solve with A	prec. MINRES, solve with $A-\sigma I$
10^{-8}	286	24	77
10^{-10}	314	28	86
10^{-12}	345	32	95

Table 3.14: Cost of a linear solve with MINRES, Test 3.8

LOBPCG outperforms all variations of inexact inverse iteration.

The results, tabulated in Table 3.13, indicate that for the targeted accuracy of 10^{-10} inexact inverse iteration, especially PInvit, is competitive. However the results for the accuracy 10^{-12} need to be handled with care, as $10^{-12} < eps\ |\lambda_n|\ /\ |\lambda_1| \approx 2.1 \times 10^{-12}$. Most runs were stopped after detecting that the targeted accuracy could not be attained. Here we report for RQIf, RQId, PInvit, and PInvit+ the number of iterations until this failure is detected. However in all cases $t^i < 2 \times 10^{-14}$ and $\|\mathbf{r}^i\| < 4 \times 10^{-8}$. For LOBPCG we report the number of iterations until $\|\mathbf{r}^i\| < 4 \times 10^{-8}$ was achieved. The small difference in the performance between LOBPCG and inexact inverse iteration is encouraging.

Test 3.8 *Consider 'bcsstk09' and solve the linear system* $A\mathbf{x} = \mathbf{b}$ *with* $\|\mathbf{b}\|_2 = 1$ *up to the accuracies* 10^{-8}, 10^{-10}, *and* 10^{-12}, *using unpreconditioned MINRES. Then repeat with preconditioned MINRES. As a preconditioner we use an incomplete Cholesky factorisation of the matrix A. To calculate the preconditioner we use the MatLab routine* cholinc *with droptol* $= 0.01$. *Finally solve the system* $(A-\sigma I)\mathbf{x} = \mathbf{b}$ *with preconditioned MINRES, where* $\sigma = \frac{1}{2}(\lambda_{20} + \lambda_{19})$. *The performance is given in Table 3.14.*

Comparing the results of Test 3.14 with all previous tests we see that solving the eigenvalue problem is not much more expensive as solving a linear system. Specially when comparing SE, PInvit and PInvit+ with the linear solve, Tables 3.6, 3.7 and 3.8 with Table 3.14, we see that solving the eigenvalue problem is about twice expensive than solving the linear system. This has been confirmed also for other starting vectors and other aimed accuracies.

3.8.4 Conclusion

In this chapter we analysed the efficiency of inexact inverse iteration using MINRES. For this we defined appropriate measures for the cost. That are the number of inner iterations per outer iteration $\mathcal{L}^i$ and the total number of inner iterations $\mathcal{T}$. In Sections 3.3 and 3.4 we provided a posteriori bounds for $\mathcal{L}^i$ and $\mathcal{T}$ for the case that inexact inverse iterations is used with unpreconditioned and preconditioned MINRES. The a posteriori bound for $\mathcal{L}^i$ links the cost of a linear solve with the progress the linear solve achieved in one outer iteration. Based on the a posteriori bound for $\mathcal{T}$ we showed that

it is beneficial to reduce the number of outer iterations, which, in practise, can be be achieved by using the RQ as shift.

Another important part of this chapter was the study of some variation of inexact inverse iteration. We proved convergence of the inverse correction method from Rüde and Schmid (1995) by linking it to inexact inverse iteration. An efficiency result was proved in Section 3.5. Further we provided a convergence proof for the algorithm proposed by Simoncini and Eldén (2002). This approach was extended to the use of any positive definite preconditioner. Based on the efficiency result we showed why this approach is superior to other approaches.

Finally we compared the studied methods using numerical examples. These examples revealed that the approach from Simoncini and Eldén (2002), SE, and our variation, PInvit, are efficient. Further we have seen that PInvit+ which applicable for any preconditioner is competitive and robust. In contrast, Test 3.5 showed that SE relies on a good choice of the stopping parameter τ_{SE} which is not a priory known. A comparison with LOBPCG and a linear solve as benchmark revealed that these eigenvalue solvers are very efficient.

Chapter 4

Convergence of Inexact Inverse Iteration for the generalised eigenvalue problem

In this chapter we consider the generalised unsymmetric eigenvalue problem, (GEP),

$$A\mathbf{x} = \lambda M\mathbf{x}, \tag{4.1}$$

with $A, M \in \mathbb{R}^{n\times n}$ where the eigenpair $(\lambda, \mathbf{x})$ is sought. Later we might refer to (4.1) as the right eigenvalue problem. In this and the following Chapters we restrict to the case where M is symmetric positive definite (spd).

As in Chapters 2 and 3 we mean by inexact inverse iteration that the linear systems arising in inverse iteration are solved only approximately. In this chapter we require only that the linear solves satisfy a residual constraint. The resulting analysis is therefore independent of the linear solver.

We start by discussing a few basic properties of the GEP in Section 4.1. There we discuss the eigen-decomposition of the matrix pair A, M, and the generalised tangent which we use as a measure for convergence. Further we derive bounds which get frequently used in later sections. This includes bounds relating the eigenvalue residual and the Rayleigh quotient with the generalised tangent.

Then in Section 4.2 we present a general convergence result for inexact inverse iteration. This convergence result, Theorem 4.2, is a key result for the remainder of this chapter and Chapter 6. As we only use a constraint condition on the residual of the linear solves, the result is independent of the method applied to obtain the next iterate, and hence it can be applied to a variety of practical variations of inexact inverse iteration.

In Section 4.3 we use this general convergence result to deduce convergence for a few practical methods. The selection of methods includes inexact inverse iteration using a

fixed shift and a decreasing tolerance. We also discuss two variations of the Rayleigh quotient iteration. Additionally we consider an update technique for the shift proposed in Wilkinson (1965). Again we consider the approach from Rüde and Schmid (1995). Further we show how the approach of Golub and Ye (2000) relates to inexact inverse iteration. In Chapter 3 we discussed the approach of Simoncini and Eldén (2002) for the symmetric eigenvalue problem, here we extend their approach to the GEP.

The above approaches can be classified as one sided approaches, as they only solve the right eigenvalue problem, (4.1). In contrast, two sided approaches solve the left eigenvalue problem

$$\mathbf{x}^H A = \lambda \mathbf{x}^H M, \tag{4.2}$$

in addition to the right eigenvalue problem. While for A symmetric and M spd, the right eigenvalue problem and the left eigenvalue problem have the same solution, now for A unsymmetric the eigenvalues are still the same but the eigenvectors differ. The advantage of using a two sided approach is that one can use the generalised RQ as shift, which provides a better, i.e. higher order, approximation of the sought eigenvalue than the standard RQ does. We discuss two practical methods based on the two sided approach in Section 4.4.

Finally in Section 4.5 we provide some numerical examples.

4.1 Some basic results

4.1.1 Jordan decomposition

In contrast to the standard symmetric eigenvalue problem for arbitrary matrix pairs A, M the GEP might not have a full set of independent solutions. However, if M is spd then a full set of *generalised eigenvectors* exists.

As we simultaneously use the solutions of the right and the left eigenvalue problem, we use super indices L and R and write

$$A\mathbf{v}_j^R = \lambda_j M \mathbf{v}_j^R \quad \text{and} \quad (\mathbf{v}_j^L)^H A = \lambda_j (\mathbf{v}_j^L)^H M. \tag{4.3}$$

Next we want to decompose the left, (4.2) and the right eigenvalue problem (4.1). For this we use the existence of a Jordan decomposition for any unsymmetric matrix B, see, for example, Golub and van Loan (1996, Theorem 7.1.9). Let $B \in \mathbb{C}^{n\times n}$ then there exists $W \in \mathbb{C}^{n\times n}$ and $J \in \mathbb{C}^{n\times n}$ such that

$$B = WJW^{-1}, \tag{4.4}$$

where $J = \text{diag}(J_j)$ is called Jordan matrix and the J_j's are referred to as Jordan

blocks. Each Jordan block has the format

$$J_j := \begin{pmatrix} \lambda_j & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_j \end{pmatrix}.$$

This definition for the Jordan blocks is common however the scaling of the upper diagonal elements is arbitrary, meaning any other value except zero would do. Further the sizes m_j of the Jordan blocks $J_j \in \mathbb{C}^{m_j \times m_j}$ add up to n.

As M is spd we can factorise $M = M^{\frac{1}{2}} M^{\frac{1}{2}}$, where $M^{\frac{1}{2}}$ is again spd. Now let (4.4) be the Jordan decomposition of $B = M^{-\frac{1}{2}} A M^{-\frac{1}{2}}$ then set $V_R := M^{-\frac{1}{2}} W$ to obtain

$$\begin{aligned} M^{-\frac{1}{2}} A M^{-\frac{1}{2}} &= W J W^{-1} \\ \Leftrightarrow \quad A M^{-\frac{1}{2}} W &= M M^{-\frac{1}{2}} W J \\ \Leftrightarrow \quad A V_R &= M V_R J. \end{aligned} \tag{4.5}$$

This is an eigenvalue decomposition of the GEP $A\mathbf{x} = \lambda M\mathbf{x}$, where M is spd. We refer to (4.5) as the Jordan decomposition. For more general eigen-decompositions, valid for any GEP see for example Turnbull and Aitken (1932) or Gantmacher (1959a,b). Similarly with $V_L^H := W^{-1} M^{-\frac{1}{2}}$ we gain for the left eigenvalue problem (4.2)

$$V_L^H A = J V_L^H M. \tag{4.6}$$

Using the Jordan decomposition (4.5) and assuming σ is not an eigenvalue of the matrix pair A, M we can write

$$\begin{aligned} (A - \sigma M) V_R &= M V_R (J - \sigma I) \\ \Leftrightarrow \quad V_R (J - \sigma I)^{-1} &= (A - \sigma M)^{-1} M V_R. \end{aligned} \tag{4.7}$$

Similar we can use (4.6) to obtain

$$\begin{aligned} V_L^H (A - \sigma M) &= (J - \sigma I) V_L^H M \\ \Leftrightarrow \quad (J - \sigma I)^{-1} V_L^H &= V_L^H M (A - \sigma M)^{-1}. \end{aligned} \tag{4.8}$$

Additionally we observe that V_L is M-orthogonal to V_R,

$$V_L^H M V_R = W^{-1} M^{-\frac{1}{2}} M M^{-\frac{1}{2}} W = W^{-1} W = I.$$

So the scaling of the left eigenvectors is implicitly given by the scaling of the right eigenvectors. The M-orthogonality of V_L to V_R implies $I = V_R^H M V_L = V_R V_L^H M = M V_R V_L^H$.

Later in the convergence analysis we like to use some bounds on the norm of a Jordan matrix, say $\widetilde{J}$. In contrast to the standard symmetric eigenvalue problem the eigenvalues of an unsymmetric matrix B no longer determine $\|B\|_2$, but the singular values do. So in order to bound $\|\widetilde{J}\|_2$ we use a bound on the singular values of the Jordan block J_j. The singular values of an unsymmetric matrix B are the square roots of the eigenvalues of the matrix BB^H which equal the eigenvalues of B^HB. Using the Theorem of Gershgorin, see for example Golub and van Loan (1996, p. 320), we can derive an appropriate bound. Let the matrix $B \in \mathbb{C}^{n\times n}$ have the elements (b_{ij}), then the Theorem of Gershgorin states that for each eigenvalues μ of B there exists a diagonal element b_{ii} such that

$$|\mu - b_{ii}| \leq \sum_{j\neq i} |b_{ij}| .$$

In case of a Jordan block matrix J with size larger than one, this reduces to three inequalities for the singular values ν of J. Denote the diagonal elements of the Jordan block by a, then the Theorem of Gershgorin provides the three inequalities

$$\begin{aligned} |\nu^2 - |a|^2 - 1| &\leq 2\,|a| \quad \text{or} \\ |\nu^2 - |a|^2 - 1| &\leq |a| \quad \text{or} \\ |\nu^2 - |a|^2\,| &\leq |a| . \end{aligned}$$

If ν is a singular value of J then at least one of the three inequalities must hold. As the first inequality gives the largest inclusion interval containing the other two, we use the first inequality to obtain a bound on the singular values. Expanding the modulus of the left-hand side of the first inequality gives $(|a| - 1)^2 \leq \nu^2 \leq (|a| + 1)^2$. However if $a \neq 0$ then $\nu > 0$, as the Jordan block is non-singular and hence its singular value decomposition is also non-singular. In case the size of the Jordan block is one, the corresponding singular value equals the modulus of the eigenvalue, i.e. $\nu = |a|$. Using the fact that the set of singular values of a block diagonal matrix, for example a Jordan matrix, is given by the union of the singular values of the diagonal blocks, we obtain the following result.

Lemma 4.1 *Given a non-singular Jordan matrix J with diagonal entries $a_1, \ldots, a_k$ then*

$$\|J\| \leq \max_j \{|a_j| + d_j\}$$

where $d_j = 1$ if a_j belongs to a Jordan block of size larger than one, otherwise $d_j = 0$. Now, if the first block is of size one with diagonal entry a_1, and all other diagonal

entries with $d_j = 1$ *have* $|a_j| > 1$ *then*

$$\|J^{-1}(I - \mathbf{e}_1\mathbf{e}_1^T)\| \quad \leq \quad (\min_{j\geq 2}\{|a_j| - d_j\})^{-1}$$

Proof: The proof for the first part follows from the above said. The second part uses the effect that the singular values of J^{-1} are the multiplicative inverse of the singular values of J. □

4.1.2 Generalised Tangent

In order to analyse the convergence of inexact inverse iteration we use the following splitting

$$\mathbf{x}^i \quad = \quad \alpha^i(c^i\mathbf{v}_1^R + s^i\mathbf{u}^i), \tag{4.9}$$

where $\mathbf{u}^i \in \text{span}(\mathbf{v}_2^R, \ldots, \mathbf{v}_n^R)$ and $\| V_L^H M\mathbf{u}^i \|_2 = 1$. Defining $\alpha^i := \| V_L^H M\mathbf{x}^i \|_2$ gives $|s^i|^2 + |c^i|^2 = 1$, as

$$\begin{aligned} 1 \quad &= \quad \frac{\|V_L^H M\mathbf{x}^i\|_2^2}{(\alpha^i)^2} \quad = \quad \|V_L^H M(c^i\mathbf{v}_1^R + s^i\mathbf{u}^i)\|_2^2 \\ &= \quad \|V_L^H M V_R \left(c^i\mathbf{e}_1 + s^i\mathbf{z}^i\right)\|_2^2 \quad = \quad \| \left(c^i\mathbf{e}_1 + s^i\mathbf{z}^i\right)\|_2^2 \quad = \quad |c^i|^2 + |s^i|^2, \end{aligned}$$

where $\mathbf{z}^i$ is implicitly defined by $V_R\mathbf{z}^i = \mathbf{u}^i$ hence $\mathbf{z}^i \perp \mathbf{e}_1$ and $\|\mathbf{z}^i\|_2 = 1$. Like in the standard symmetric eigenvalue problem we interpret s^i as a generalised sine and c^i as a generalised cosine. As a measure for convergence we define the generalised tangent

$$t^i := \frac{\|V_L^H M(I - \mathbf{v}_1^R\mathbf{v}_1^L M)\mathbf{x}^i\|}{|\mathbf{v}_1^L M\mathbf{x}^i|} = \frac{\alpha^i\,|s^i|}{\alpha^i\,|c^i|} = \frac{|s^i|}{|c^i|} \tag{4.10}$$

which is independent of the scaling α^i.

To bound the eigenvalue residual or the distance between the RQ and the sought eigenvalue we use a generalisation of the *numerical radius*. Given a matrix A we define the numerical radius as

$$\mathcal{R}(A) \quad := \quad \max_{\mathbf{z}\neq 0} \frac{|\mathbf{z}^H A\mathbf{z}|}{|\mathbf{z}^H\mathbf{z}|}, \tag{4.11}$$

(see, for example, Ipsen (1998b)). We extend this definition to the *generalised numerical radius*

$$\mathcal{R}_G(\sigma) \quad := \quad \max_{\mathbf{z},\mathbf{u}\neq 0} \frac{|\mathbf{z}^H(A - \sigma M)\mathbf{u}|}{\|V_L^H M\mathbf{z}\|\,\|V_L^H M\mathbf{u}\|}.$$

Further we define

$$\widetilde{\mathcal{R}} \;:=\; \mathcal{R}_G(\lambda_1)\max_{\mathbf{z}\neq 0}\frac{\|V_L^H M\mathbf{z}\|^2}{\|\mathbf{z}\|_M^2}. \tag{4.12}$$

Now we consider the Rayleigh quotient (RQ), ϱ, for a given vector $\mathbf{z}\in\mathbb{C}^n$,

$$\varrho \;:=\; \frac{\mathbf{z}^H A\mathbf{z}}{\mathbf{z}^H M\mathbf{z}}. \tag{4.13}$$

Then for any non-zero vector $\mathbf{x}$ with splitting $\mathbf{x}=\alpha(c\mathbf{v}_1^R+s\mathbf{u})$ according to (4.9) we obtain

$$\begin{aligned}|\varrho-\lambda_1| &= \frac{|\mathbf{x}^H A\mathbf{x}-\lambda_1\mathbf{x}^H M\mathbf{x}|}{\mathbf{x}^H M\mathbf{x}}\\ &= |s|\,\frac{|\mathbf{x}^H(A-\lambda_1 M)\mathbf{u}|}{\|V_L^H M\mathbf{x}\|}\,\frac{\|V_L^H M\mathbf{x}\|^2}{\mathbf{x}^H M\mathbf{x}} \;\le\; |s|\,\widetilde{\mathcal{R}},\end{aligned} \tag{4.14}$$

where we used that $\|V_L^H M\mathbf{u}\|=1$.

Later we need the fact that $\|J-\lambda_1 I\|\le\widetilde{\mathcal{R}}$. The bound is not obvious, as J is given by the spectrum of the eigenvalue problem $A\mathbf{x}=\lambda M\mathbf{x}$, while $\widetilde{\mathcal{R}}$ is related to the matrix $A-\lambda_1 M$. However, using the eigen decomposition we obtain

$$\begin{aligned}\|J-\lambda_1 I\| &= \|V_L^H(A-\lambda_1 M)V_R\|\\ &\le \max_{\mathbf{x},\mathbf{u}\neq\mathbf{0}}\frac{|\mathbf{x}^H V_L^H(A-\lambda_1 M)V_R\mathbf{u}|}{\|\mathbf{x}\|\,\|\mathbf{u}\|}\\ &= \max_{\mathbf{x},\mathbf{u}\neq\mathbf{0}}\frac{|\mathbf{x}^H(A-\lambda_1 M)\mathbf{u}|}{\|\mathbf{x}^H MV_R\|\,\|V_L^H M\mathbf{u}\|}\\ &= \max_{\mathbf{x},\mathbf{u}\neq\mathbf{0}}\frac{|\mathbf{x}^H(A-\lambda_1 M)\mathbf{u}|}{\|V_L^H M\mathbf{x}\|\,\|V_L^H M\mathbf{u}\|}\,\frac{\|V_L^H M\mathbf{x}\|}{\|\mathbf{x}^H MV_R\|}\\ &\le \mathcal{R}_G(\lambda_1)\max_{\mathbf{x},\mathbf{u}\neq\mathbf{0}}\frac{\|V_L^H M\mathbf{x}\|}{\|\mathbf{x}^H MV_R\|}\\ &\le \widetilde{\mathcal{R}}.\end{aligned} \tag{4.15}$$

For the last inequality we used

$$\begin{aligned}|\mathbf{x}^H M\mathbf{x}| &= |\mathbf{x}^H MV_R V_L^H M\mathbf{x}|\\ &\le \|\mathbf{x}^H MV_R\|\,\|V_L^H M\mathbf{x}\|\,.\end{aligned}$$

For the standard symmetric eigenvalue problem the RQ minimises the 2-norm of the residual $\widetilde{\mathbf{r}}(\mu):=A\mathbf{z}-\mu\mathbf{z}$ for a given vector $\mathbf{z}$. However this is no longer the case for the GEP, where the RQ is the minimiser of $\|\mathbf{r}(\mu)\|_{M^{-1}}$, with $\mathbf{r}(\mu):=A\mathbf{z}-\mu M\mathbf{z}$, while the minimiser for $\|\mathbf{r}(\mu)\|_2$ is given by $\mu^*:=\mathbf{z}^H MA\mathbf{z}/(\mathbf{z}^H MM\mathbf{z})$. For bounding

the eigenvalue residual we use that for some constant $C_{12} > 0$ the change of norms can be bounded, $\max_{\mathbf{z}\neq 0} \|\mathbf{z}\| \ / \ \|\mathbf{z}\|_{M^{-1}} \leq C_{12}$. Then together with the optimality of the RQ we gain for $\mathbf{x} = \alpha(c\mathbf{v}_1^R + s\mathbf{u})$ and $\mathbf{r} := \mathbf{r}(\varrho)$ that

$$\begin{aligned} \|(A-\varrho M)\mathbf{x}\|_2 \ = \ \|\mathbf{r}\|_2 \ &\leq \ C_{12}\,\|\mathbf{r}\|_{M^{-1}} \leq \|(A-\lambda_1 M)\mathbf{x}\|_{M^{-1}} \\ &\leq \ \alpha\,|s|\,C_{12}\,\|(A-\lambda_1 M)\mathbf{u}\|_{M^{-1}} \\ &\leq \ \alpha\,|s|\,C_{12}\,\|(M^{-\frac{1}{2}}AM^{-\frac{1}{2}}-\lambda_1 I)M^{\frac{1}{2}}V_R\mathbf{z}\| \\ &\leq \ \alpha\,|s|\,C_{12}\,\|J-\lambda_1 I\|\|M^{\frac{1}{2}}V_R\|\,. \end{aligned} \tag{4.16}$$

For our convenience we define

$$\mathcal{R}^* \ := \ C_{12}\,\|J-\lambda_1 I\|\|M^{\frac{1}{2}}V_R\|, \tag{4.17}$$

and obtain $\|\mathbf{r}\|_2 \leq \alpha\mathcal{R}^*\,|s|$.

4.2 Convergence of inexact inverse iteration

In this section we present our general convergence result for inexact inverse iteration, as given in Algorithm 5. In the algorithm we have not specified how we update $\mathbf{x}^{i+1}$. In practise many different techniques exist, so in order not to restrict to a specific one we consider the update

$$\mathbf{x}^{i+1} \ = \ \varphi(\mathbf{y}^i)\mathbf{y}^i, \tag{4.18}$$

where φ is a scalar function. Typical choices for φ are $\varphi(\mathbf{y}^i) = \|\mathbf{y}^i\|_M^{-1}$ and $\varphi(\mathbf{y}^i) = (\mathbf{z}^H\mathbf{y}^i)^{-1}$ for some fixed $\mathbf{z}$. Note, in Algorithm 5, that the linear solve step uses a general right-hand side $\mathbf{b}^i$. Standardly the right-hand side is chosen to be $\mathbf{b}^i = M\mathbf{x}^i$. However, later in Section 4.3.4 we consider a modified right-hand side $\mathbf{b}^i = P\mathbf{x}^i$ for some matrix P.

In order to proceed with the convergence analysis we assume that the sought eigenvalue, say λ_1, is simple and well separated. To make this statement more precise we define

$$gap \ := \ \min_{j\geq 2}\{|\lambda_j-\lambda_1| - d_j\} \tag{4.19}$$

where $d_j = 1$ if λ_j is defective and $d_j = 0$ otherwise. From now on we assume that $gap > 0$ and σ is such that

$$0 \ < \ |\lambda_1-\sigma| \ < \ \frac{1}{2}gap. \tag{4.20}$$

Algorithm 5: Inexact Inverse Iteration for GEPs

Given $\mathbf{x}^0 \neq \mathbf{0}$,
For $i = 0, 1, 2, \ldots$

- Choose σ^i, $\mathbf{b}^i$ and τ^i,
- Inexact solve $(A - \sigma^i M)\mathbf{y}^i = \mathbf{b}^i$ such that $\|\mathbf{b}^i - (A - \sigma^i M)\mathbf{y}^i\| \leq \tau^i$,
- Update $\mathbf{x}^{i+1}$ using $\mathbf{y}^i$.
- Test for convergence

As a result we obtain for $\|(J - \sigma I)^{-1}(I - \mathbf{e}_1\mathbf{e}_1^T)\|$ by using Lemma 4.1

$$\begin{aligned} \|(J - \sigma I)^{-1}(I - \mathbf{e}_1\mathbf{e}_1^T)\| &\leq (\min_{j\geq 2}\{|\lambda_j - \sigma| - d_j\})^{-1} \\ &\leq (\min_{j\geq 2}\{|\lambda_j - \lambda_1| - d_j\} - |\lambda_1 - \sigma|)^{-1} \\ &= (gap - |\lambda_1 - \sigma|)^{-1} \end{aligned} \tag{4.21}$$

By using inexact solves for the update equation $(A - \sigma^i M)\mathbf{y}^i = \mathbf{b}^i$ we obtain a residual, let this residual be defined by

$$\mathbf{res}^i := \mathbf{b}^i - (A - \sigma^i M)\mathbf{y}^i. \tag{4.22}$$

Rearranging this equation and using the scaling of $\mathbf{x}^{i+1}$, (4.18), gives

$$(A - \sigma^i M)\mathbf{x}^{i+1} = \varphi(\mathbf{y}^i)(\mathbf{b}^i - \mathbf{res}^i).$$

The assumption $0 < |\lambda_1 - \sigma^i| < \frac{1}{2}gap$ implies that $(A - \sigma^i M)$ is invertible, hence we gain the update equation

$$\mathbf{x}^{i+1} = \varphi(\mathbf{y}^i)(A - \sigma^i M)^{-1}(\mathbf{b}^i - \mathbf{res}^i). \tag{4.23}$$

We observe that $(\mathbf{v}_1^L)^H M\mathbf{x}^{i+1} = \alpha^{i+1}c^{i+1}$, and $(\mathbf{v}_1^L)^H = \mathbf{e}_1^T V_L^H$. Hence by premultiplying the update equation (4.23) by $(\mathbf{v}_1^L)^H M$ and using $V_L^H M(A - \sigma^i M)^{-1} = (J - \sigma^i I)V_L^H$, see (4.8), we gain

$$\begin{aligned} \alpha^{i+1}c^{i+1} &= \varphi(\mathbf{y}^i)\mathbf{e}_1^T V_L^H M(A - \sigma^i M)^{-1}(\mathbf{b}^i - \mathbf{res}^i) \\ &= \varphi(\mathbf{y}^i)\mathbf{e}_1^T(J - \sigma^i I)^{-1}V_L^H(\mathbf{b}^i - \mathbf{res}^i) \end{aligned}$$

$$= \varphi(\mathbf{y}^i)(\lambda_1 - \sigma^i)^{-1}(\mathbf{v}_1^L)^H(\mathbf{b}^i - \mathbf{res}^i) \tag{4.24}$$

for the cosine part.

To obtain a bound on $|s^{i+1}|$ we use the matrix

$$Q := (I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H M = V_L^H M(I - \mathbf{v}_1^R(\mathbf{v}_1^L)^H M).$$

which projects the sought eigen-directions to zero. As $V_L^H M\mathbf{v}_1^R = \mathbf{e}_1$ and $(\mathbf{v}_1^L)^H M\mathbf{u}^{i+1} = 0$ we obtain

$$\begin{aligned} Q\mathbf{x}^{i+1} &= (I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H M\mathbf{x}^{i+1} \\ &= \alpha^{i+1}s^{i+1}(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H M\mathbf{u}^{i+1} \\ &= \alpha^{i+1}s^{i+1}V_L^H M\mathbf{u}^{i+1}. \end{aligned} \tag{4.25}$$

Again we use that $V_L^H M(A - \sigma^i M)^{-1} = (J - \sigma^i I)^{-1}V_L^H$, see (4.8), therefore we obtain using the update equation

$$\begin{aligned} Q\mathbf{x}^{i+1} &= Q\varphi(\mathbf{y}^i)(A - \sigma^i M)^{-1}(\mathbf{b}^i - \mathbf{res}^i) \\ &= \varphi(\mathbf{y}^i)(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H M(A - \sigma^i M)^{-1}(\mathbf{b}^i - \mathbf{res}^i) \\ &= \varphi(\mathbf{y}^i)(I - \mathbf{e}_1\mathbf{e}_1^T)(J - \sigma^i I)^{-1}V_L^H(\mathbf{b}^i - \mathbf{res}^i) \\ &= \varphi(\mathbf{y}^i)(J - \sigma^i I)^{-1}(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H(\mathbf{b}^i - \mathbf{res}^i). \end{aligned} \tag{4.26}$$

Now we combine the two equations for $Q\mathbf{x}^{i+1}$, (4.25) and (4.26), and take norms to obtain by using (4.21)

$$\begin{aligned} \alpha^{i+1}\,|s^{i+1}| &= \|V_L^H M\mathbf{u}^{i+1}\alpha^{i+1}s^{i+1}\| \\ &= \|\varphi(\mathbf{y}^i)(J - \sigma^i I)^{-1}(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H(\mathbf{b}^i - \mathbf{res}^i)\| \\ &\leq |\varphi(\mathbf{y}^i)|\|(J - \sigma^i I)^{-1}(I - \mathbf{e}_1\mathbf{e}_1^T)\|\,\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H(\mathbf{b}^i - \mathbf{res}^i)\| \\ &\leq \frac{|\varphi(\mathbf{y}^i)|}{gap - |\lambda_1 - \sigma^i|}\Big(\,\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\| + \|V_L^H\mathbf{res}^i\|\,\Big). \end{aligned} \tag{4.27}$$

Finally we assume that the inexact solve is such that $|(\mathbf{v}_1^L)^H\mathbf{b}^i| > \|V_L^H\mathbf{res}^i\|$ holds. Now we divide (4.27) by the modulus of (4.24) to obtain the one-step bound

$$t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{gap - |\lambda_1 - \sigma^i|}\,\frac{\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\| + \|V_L^H\mathbf{res}^i\|}{|(\mathbf{v}_1^L)^H\mathbf{b}^i| - \|V_L^H\mathbf{res}^i\|}. \tag{4.28}$$

This one step-bound plays here a similar role as the one-step bound (2.18) in case of the standard symmetric eigenvalue problem. Obviously convergence is achieved if one of the two terms on the right-hand side in (4.28) is bounded and the other tends to zero. Similarly, higher order convergence is achieved if both terms tend to zero or

the first term tends towards zero superlinearly. In the following theorem we state the convergence of Algorithm 5 more precisely.

Theorem 4.2 *Given $A, M \in \mathbb{R}^{n\times n}$ with M spd. Let the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ have the simple eigenvalue λ_1 with gap > 0. Assume $\exists C_1, C_2, C_3$ and $\beta \in \mathbb{R}^+$ and $\gamma_1, \gamma_2 \in [0,1]$ such that for $\beta + \gamma \geq 1$ where $\gamma := \min\{\gamma_1, \gamma_2\}$ the conditions*

a) $0 < |\lambda_1 - \sigma^i| \leq \min\{\frac{1}{4}gap, C_1(t^i)^\beta\}$,

b) $\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\|_2 \leq C_2\,|s^i|^{\gamma_1}|(\mathbf{v}_1^L)^H\mathbf{b}^i|$,

c) $\|V_L^H\mathbf{res}^i\|_2 \leq C_3\,|s^i|^{\gamma_2}|(\mathbf{v}_1^L)^H\mathbf{b}^i|$ *with* $C_3\,|s^0|^{\gamma_2} < 1$

hold. If the initial approximation $\mathbf{x}^0$ satisfies

$$q := \frac{4C_1(t^0)^{\beta+\gamma-1}}{3gap}\;\frac{C_2\,|s^0|^{\gamma_1-\gamma} + C_3\,|s^0|^{\gamma_2-\gamma}}{1 - C_3\,|s^0|^{\gamma_2}} < 1 \tag{4.29}$$

then $t^{i+1} \leq qt^i$ and $span\{\mathbf{x}^i\} \to span\{\mathbf{v}_1^R\}$ while $\varrho^i \to \lambda_1$.

Proof: We use induction to show $t^{i+1} \leq t^i q$ for q given in (4.29) which implies $|s^{i+1}| \leq |s^0|$ and $|c^i| \geq |c^0|$. Starting with the one-step bound by using first condition a), then b) and c) we get

$$\begin{aligned} t^{i+1} &\leq \frac{4C_1\,|t^i|^\beta}{3gap}\;\frac{\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\| + \|V_L^H\mathbf{res}^i\|}{|(\mathbf{v}_1^L)^H\mathbf{b}^i| - \|V_L^H\mathbf{res}^i\|} \\ &\leq \frac{4C_1\,|t^i|^\beta}{3gap}\;\frac{C_2\,|s^i|^{\gamma_1} + C_3\,|s^i|^{\gamma_2}}{1 - C_3\,|s^i|^{\gamma_2}} \\ &\leq t^i\,|t^i|^{\beta+\gamma-1}\,\frac{4C_1}{3gap}\;\frac{C_2\,|s^i|^{\gamma_1-\gamma} + C_3\,|s^i|^{\gamma_2-\gamma}}{1 - C_3\,|s^i|^{\gamma_2}} \leq t^i q. \end{aligned} \tag{4.30}$$

Then by the condition on the initial approximation we have $q^i \leq q$ and thereby $t^{i+1} \leq qt^i \leq (q)^{i+1}t^0$ with $q < 1$, hence $t^i \to 0$. As $t^i \to 0$ so $s^i \to 0$ and hence $|\varrho^i - \lambda_1| \to 0$ and $\text{span}\{\mathbf{x}^i\} \to \text{span}\{\mathbf{v}_1^R\}$. □

We now present a corollary tuned to the standard right-hand side $\mathbf{b}^i = M\mathbf{x}^i$.

Corollary 4.3 *Given $A, M \in \mathbb{R}^{n\times n}$ with M spd. Let the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ have the simple eigenvalue λ_1 with gap > 0. Further, in Algorithm 5 let $\mathbf{b}^i = M\mathbf{x}^i$. Assume $\exists C_1, C_3$ and $\beta \in \mathbb{R}^+$ and $\gamma_1, \gamma_2 \in [0,1]$ such that for $C_2 = 1/\,|c^0|$ and $\beta + \gamma \geq 1$ the conditions*

a) $0 < |\lambda_1 - \sigma^i| \leq \min\{\frac{1}{4}gap, C_1(t^i)^\beta\}$,

b) $\|V_L^H\mathbf{res}^i\|_2 \leq C_3\,|s^i|^{\gamma}|c^0|$

hold. If the initial approximation $\mathbf{x}^0$ *satisfies*

$$q := \frac{4C_1(t^0)^{\beta+\gamma-1}}{3gap} \frac{C_2 |s^0|^{1-\gamma} + C_3}{1 - C_3 |s^0|^{\gamma}} < 1 \tag{4.31}$$

then $t^{i+1} \leq qt^i$ *and* $span\{\mathbf{x}^i\} \to span\{\mathbf{v}_1^R\}$ *while* $\varrho^i \to \lambda_1$.

Proof: With $\mathbf{b}^i = M\mathbf{x}^i$ condition b) reduces to $|s^i| \leq C_2 |s^i||c^i|$. Hence we set $\gamma_1 = 1$ and so $\gamma_2 = \gamma$, and thereby q as in (4.29) reduces to q as in (4.31). □

Applying the one-step bound for the GEP, (4.28), to the standard symmetric eigenvalue problem we obtain the one-step bound (3.55) as discussed in Section 3.6. In addition setting $\mathbf{b}^i = \mathbf{x}^i$ we regain the one-step bound (2.18) discussed in Chapter 2.

The conditions of Theorem 4.2 ensure that both terms on the right-hand side of (4.28) are bounded and that at least one of them tends to zero. For example, given a good enough initial guess and an appropriate residual condition the condition of Theorem 4.2 ensures convergence if the shift tends towards the desired eigenvalue. We admit that the conditions of Theorem 4.2 are non-practical, as, for example, $\mathbf{v}_1^L$ is unknown in practice.

In Theorem 4.2 we stated only $t^{i+1} \leq qt^i$ for a fixed q, however we proved the following remark.

Remark 4.4 *Under the conditions of Theorem 4.2, the rate of convergence* q^i, *so* $t^{i+1} \leq q^i t^i$, *is given by*

$$q^i := \frac{4C_1 |t^i|^{\beta+\gamma-1}}{3gap} \frac{C_2 |s^i|^{\gamma_1-\gamma} + C_3 |s^i|^{\gamma_2-\gamma}}{1 - C_3 |s^i|^{\gamma_2}}.$$

For a fixed shift, $\sigma^i = \sigma^0$, a bound similar to the one-step bound, (4.28), has been obtained by Golub and Ye (2000). We discuss their approach based on a residual equation later in Section 4.3.5. Neymeyr (2001b) considers A, M with positive real spectra, $0 < \lambda_1 < \lambda_2 < \ldots$. For λ_1 being the smallest eigenvalue he proves convergence for a method with fixed shift. The result is based on the monotonic reduction of the RQ. In Neymeyr (2002) a similar result for exact solves is presented in a way that allows the application of variable shifts. Notay (2003) also considers the case of a real positive definite eigenvalue problem. He proves higher order convergence for an inexact Rayleigh quotient iteration.

4.3 Practical methods

Our general treatment of inexact inverse iteration gives rise to various practical methods, of which we consider only a few in detail.

We start by discussing inexact inverse iteration with a fixed shift and a decreasing tolerance. Next we consider four variations of variable shifts, the first two use the Rayleigh quotient, the other two a generalisation of the RQ which we call Wilkinson update. These five methods use the standard right-hand side, later in Chapter 6 we see that using the standard right-hand side and solving the linear system with GMRES leads to high costs per solve. In the following we consider three methods which have lower cost per solve when using GMRES as linear solver. The first method we discuss is the approach from Simoncini and Eldén (2002) which we discussed for the standard symmetric eigenvalue problem in Section 3.6. The other two methods we consider are based on solving residual equations, one is the approach from Rüde and Schmid (1995) and the other the approach from Golub and Ye (2000).

Numerical results for these methods are presented in Section 4.5.

4.3.1 Fixed shift

A straight-forward approach of implementing inexact inverse iteration is to use a fixed shift $\sigma^i = \sigma^0$ and to reduce the residual tolerance τ^i as the outer iteration proceeds. There are several ways to decrease the residual constraint, such as $\tau^i = p(q)^i$ with $p > 0$ and $q \in (0,1)$ or $\tau^i \propto s^i$. Here we consider the following method using the standard right-hand side $\mathbf{b}^i = M\mathbf{x}^i$.

InvitFd is Algorithm 5 (p. 89) with

$$\sigma^i = \sigma^0, \quad \tau^i = \min\{\tau^0 \, \|V_L^H M\mathbf{x}^i\|, \widetilde{C}_3 \, \|\mathbf{r}^i\|\}, \quad \mathbf{b}^i = M\mathbf{x}^i. \tag{4.32}$$

The following result states the linear convergence of this approach, its proof is technical and presented here only for completeness.

Corollary 4.5 *Apply InvitFd, that is Algorithm 5 with (4.32) to the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ with $A, M \in \mathbb{R}^{n\times n}$, where M is spd. Let the shift satisfy $0 < |\lambda_1 - \sigma^0| \leq \frac{1}{4}gap$, and the initial guess satisfy $t^0 \leq \frac{1}{4}$ while in (4.32) assume $\widetilde{C}_3 \leq (\|V_L^H\| \, \mathcal{R}^*)^{-1}$, then $t^i \to 0$ linearly.*

Proof: The condition $t^0 \leq \frac{1}{4}$ implies that

$$|c^0| = \sqrt{1 - |s^0|^2} \geq \sqrt{1 - |t^0|^2} = \sqrt{\frac{15}{16}} \geq \frac{15}{16}.$$

We now apply Theorem 4.2 with $C_1 = \frac{1}{4}gap$ and $C_2 = C_3 = \frac{16}{15}$ while $\beta = 0$ and $\gamma_1 = \gamma_2 = \gamma = 1$. Condition b) is satisfied as

$$\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\| = |s^i| \, \alpha^i \leq |s^i| \, \frac{16}{15} \, |c^i| = |s^i| \, \frac{16}{15} \, |(\mathbf{v}_1^L)^H\mathbf{b}^i|,$$

where we use that $|c^i| > \frac{15}{16}$, similar to Corollary 4.6. Due to the bound on the eigenvalue residual (4.16) we obtain for condition c)

$$\begin{aligned} \|V_L^H \mathbf{res}^i\| &\leq \|V_L^H\| \|\mathbf{res}^i\| \leq \|V_L^H\| \, \tau^i \leq \|V_L^H\| \, \widetilde{C}_3 \, \|\mathbf{r}^i\| \\ &\leq \|V_L^H\| \, \widetilde{C}_3 \mathcal{R}^* \alpha^i \, |s^i| \leq |s^i| \, \frac{16}{15} \, |(\mathbf{v}_1^L)^H \mathbf{b}^i| \, . \end{aligned}$$

As $t^0 \leq \frac{1}{4}$ we have also $C_3 \, |s^0| < 1$.

Then we obtain for q as defined in Theorem 4.2

$$q \leq \frac{4C_1}{3gap} \, \frac{C_2 + C_3}{1 - C_3 \, |s^0|} \leq \frac{1}{3} \, \frac{\frac{16}{15} + \frac{16}{15}}{1 - \frac{16}{15} \, \frac{1}{4}} = \frac{32}{33}.$$

Hence we can apply Theorem 4.2 and obtain the claimed convergence. □

So using the standard right-hand side $\mathbf{b}^i = M\mathbf{x}^i$ and a fixed shift we obtain linear convergence if the residual constraint is reduced when the outer iteration proceeds. However the rate of convergence might be less than with exact solves depending on $\widetilde{C}_3$. To make this more precise we use the one-step bound (4.28) and assume $\|V^L \mathbf{res}^i\| \ll c^i$, then $t^{i+1} \leq q t^i$ with $q \approx q_0(1 + \widetilde{C}_3 \, \|\mathbf{r}^i\| \, / \, |s^i|)$, where q_0 is the convergence rate for exact solves. So increasing $\widetilde{C}_3$ in (4.32) leads to a slow down in the convergence and eventually to a lack of convergence. As $\|\mathbf{r}^i\|$ is linear in $|s^i|$ there exists a positive constant, say C_7 such that $C_7 \, \|\mathbf{r}^i\| \leq |s^i|$. Now chosing $\widetilde{C}_3$ such that $\widetilde{C}_3 \ll C_7$ does not improve the outer convergence compared with $\widetilde{C}_3 = C_7$. However choosing $\widetilde{C}_3 \ll C_7$ might lead to more difficult linear solves.

Later in Section 4.5, see Test 4.1, we provide numerical results illustrating the convergence of InvitFd.

4.3.2 Rayleigh quotient

Similar to the standard eigenvalue problem we consider two variations of the RQI. The first method is the Rayleigh quotient iteration with fixed tolerance and standard right-hand side.

RQIf is Algorithm 5 (p. 89) with

$$\sigma^i = \varrho^i, \quad \tau^i = \tau^0 \, \|V_L^H M \mathbf{x}^i\|, \quad \mathbf{b}^i = M\mathbf{x}^i. \tag{4.33}$$

The second method is the Rayleigh quotient iteration with decreasing tolerance and standard right-hand side.

RQId is Algorithm 5 (p. 89) with

$$\sigma^i = \varrho^i, \quad \tau^i = \min\{\tau^0 \, \|V_L^H M\mathbf{x}^i\|, \widetilde{C}_3 \, \|\mathbf{r}^i\|\}, \quad \mathbf{b}^i = M\mathbf{x}^i. \tag{4.34}$$

The following result uses $\widetilde{\mathcal{R}}$ as defined in (4.12), which is used to bound the error in the RQ, $|\varrho^i - \lambda_1| \leq |s^i| \, \widetilde{\mathcal{R}}$, see (4.14). Again the result is just technical and presented only for completeness.

Corollary 4.6 *Apply RQIf, that is Algorithm 5 with (4.33) to the GEP* $A\mathbf{x} = \lambda M\mathbf{x}$ *with* $A, M \in \mathbb{R}^{n\times n}$, *where* M *is spd. Let the conditions*

$$t^0 \leq \frac{gap}{4\widetilde{\mathcal{R}}} \quad \textit{and} \quad \tau^0 \leq \frac{3gap}{32 \, \|V_L\| \, \widetilde{\mathcal{R}}},$$

be satisfied then $t^i \to 0$ *(at least) linearly.*

Proof: Using $|\varrho^i - \lambda_1| \leq |s^i| \, \widetilde{\mathcal{R}}$, see (4.14) we set $C_1 = \widetilde{\mathcal{R}}$. Further we use Theorem 4.2 with $\beta = \gamma_1 = 1$ and $\gamma_2 = \gamma = 0$ while $C_2 = \frac{16}{15}$ and $C_3 = gap(10\widetilde{\mathcal{R}})^{-1}$. For condition b) we observe that

$$\begin{aligned} \|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\| &= \|\alpha^i s^i V_L^H M\mathbf{u}^i\| = \alpha^i \, |s^i|, \text{ and} \\ |(\mathbf{v}_1^L)^H\mathbf{b}^i| &= |(\mathbf{v}_1)^H M\mathbf{x}^i| = \alpha^i \, |c^i| \, . \end{aligned}$$

With $|s^i| \leq s^0 \leq t^0 \leq gap(4\widetilde{\mathcal{R}})^{-1} \leq \frac{1}{4}$ follows that $\frac{15}{16} < \sqrt{\frac{15}{16}} \leq \sqrt{1 - |s^0|} \leq |c^0| \leq |c^i|$, hence condition b) is satisfied. Next we observe

$$\begin{aligned} \|V_L^H\mathbf{res}^i\| &\leq \|V_L\| \|\mathbf{res}^i\| \leq \|V_L\| \, \tau^0 \, \|V_L^H M\mathbf{x}^i\| \leq \frac{3gap}{32\widetilde{\mathcal{R}}}\alpha^i \\ &\leq \frac{3gap}{32\widetilde{\mathcal{R}}} \frac{16}{15} \, |c^i| \, \alpha^i \leq C_3 \, |c^i| \, \alpha^i. \end{aligned}$$

Hence condition c) is satisfied as $C_3 < 1$ due to $gap \leq \widetilde{\mathcal{R}}$. Finally we observe for the condition on the initial guess by using $gap < \widetilde{\mathcal{R}}$,

$$q = \frac{4\widetilde{\mathcal{R}}}{3gap} \frac{C_2 \, |s^0| + C_3}{1 - C_3} \leq \frac{1}{3} \frac{\frac{4gap}{15\widetilde{\mathcal{R}}} + \frac{1gap}{10\widetilde{\mathcal{R}}}}{\frac{9}{10}} \leq \frac{44}{81}.$$

□

In general we expect only linear convergence of RQIf. However if the tolerance is sufficiently small, then in the first few outer iterations the convergence might appear to be quadratic. Eventually the superlinearity fades away and linear convergence is attained. We use Remark 4.4 and the definition of RQIf then the rate of convergence

is given by

$$q^i = \frac{4C_1}{gap} \frac{C_2 |s^i| + C_3}{1 - C_3}.$$

Hence by reducing the tolerance, which in term gives a smaller C_3, we obtain faster convergence. Due to this effect we expect that RQIf will need fewer iterations than InvitFd, at least if the tolerance τ_0 is chosen reasonably tight.

To state the convergence result for RQId we use again $\widetilde{\mathcal{R}}$, see (4.12), and additionally $\mathcal{R}^*$ as defined in (4.17). We use $\mathcal{R}^*$ to bound $\|\mathbf{r}^i\| \leq |s^i| \, \mathcal{R}^* \, \|V_L^H M \mathbf{x}^i\|$.

Corollary 4.7 *Apply RQId, that is Algorithm 5 with (4.34), to the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ with $A, M \in \mathbb{R}^{n \times n}$, where M is spd. Let the conditions*

$$t^0 \leq \frac{gap}{4\widetilde{\mathcal{R}}} \quad \text{and} \quad \tau^0 \leq \frac{3gap}{32 \, \|V_L\| \, \widetilde{\mathcal{R}}},$$

be satisfied then $t^i \to 0$ and the convergence is locally quadratic.

Proof: As the conditions of Corollary 4.6 are satisfied the convergence is ensured and we can make use of Theorem 4.2. To prove the local rate of convergence we assume that t^i is small enough such that $\widetilde{C}_3 t^i \mathcal{R}^* \leq \tau^i$. Now we use Theorem 4.2 with $\beta = \gamma_1 = \gamma_2 = \gamma = 1$ and $C_3 = \frac{16}{15}\widetilde{C}_3 \mathcal{R}^*$ where $\mathcal{R}^*$ as defined in (4.17), while $C_2 = \frac{16}{15}$. With the one-step bound (4.28) we gain

$$t^{i+1} \leq t^i \frac{4\mathcal{R}^*}{3gap} \frac{C_2 |s^i| + C_3 |s^i|}{1 - C_3 |s^i|} \leq (t^i)^2 \frac{40\mathcal{R}^*}{27gap}(C_2 + C_3),$$

as $C_3 t^i = \frac{16}{15}\widetilde{C}_3 \mathcal{R}^* t^i \leq \frac{1}{10}$. □

Corollary 4.7 states convergence and locally quadratic convergence (i.e. there exists *const* such that for $t^i \leq const$ the convergence is quadratic), however we expect superlinear convergence from the offset, which makes this method very competitive. We use Remark 4.4 then the local rate of convergence for RQId is given by $\frac{4}{3}C_1 gap^{-1}(1 + C_3)$. This differs from the rate for exact solves only by the factor $1 + C_3$.

For both methods, RQIf and RQId, the convergence area reduces when the conditioning of the sought eigenvalues deteriorates, as both $\widetilde{\mathcal{R}}$ and $\|V_L\|$ increase. This leads one to consider other choices for the shift σ^i, for example, the Wilkinson update or the generalised RQ.

Later in Section 4.5 we supply a few numerical results illustrating the convergence, see Test 4.2.

Algorithm 6: Inexact inverse iteration using Wilkinson update

Given $\mathbf{x}^0$, σ^0 and $\mathbf{z}$,
For $i = 0, 1, 2, \ldots$

- Choose τ^i and $\mathbf{b}^i$,
- Inexact solve $(A - \sigma^i M)\mathbf{y}^i = \mathbf{b}^i$ such that $\|\mathbf{b}^i - (A - \sigma^i M)\mathbf{y}^i\| \leq \tau^i$,
- Update $\sigma^{i+1} = \sigma^i + \mathbf{z}^H(\mathbf{b}^i + \mathbf{res}^i)/(\mathbf{z}^H M\mathbf{y}^i)$,
- Update $\mathbf{x}^{i+1} = \mathbf{y}^i/(\mathbf{z}^H M\mathbf{y}^i)$,
- Test for convergence

4.3.3 Wilkinson update

The update formula $\sigma^{i+1} = \sigma^i + 1/(\mathbf{z}^H\mathbf{y}^i)$ has no specific name in the literature, however it goes back, at least, to Wilkinson (1965, Chapter 9 §10). Therefore we call it Wilkinson update, merely to distinguish this choice from the RQ and a fixed shift. We point out that this choice of shift is different to the one Parlett (1980, p.149) calls Wilkinson shift. A more detailed account on the Wilkinson shift can be found in Trefethen and Bau (1997, p. 222).

In case of the standard unsymmetric eigenvalue problem Wilkinson used the update formula $\sigma^{i+1} = \sigma^i + 1/(\mathbf{z}^H\mathbf{y}^i)$ together with the scaling $\mathbf{z}^H\mathbf{x}^i = 1$, which results in the equality $\sigma^{i+1} = (\mathbf{z}^H A\mathbf{y}^i)/(\mathbf{z}^H\mathbf{y}^i)$. Hence the update is a generalisation of the RQ. We now consider the same generalisation of the RQ for the GEP

$$\sigma^{i+1} = \frac{\mathbf{z}^H A\mathbf{y}^i}{\mathbf{z}^H M\mathbf{y}^i}, \tag{4.35}$$

which we refer to as the Wilkinson update. This update requires $\mathbf{z}^H M\mathbf{v}_1^R \neq 0$.

So far we have no bound on $|\lambda_1 - \sigma^i|$, when σ^i is obtained by the Wilkinson update. To establish $|\lambda_1 - \sigma^{i+1}| \leq C_1 |s^{i+1}|$ for some $C_1 > 0$ we use that with $\mathbf{z}^H M\mathbf{y}^i \neq 0$

$$\begin{aligned} \lambda_1 - \sigma^{i+1} &= \lambda_1 \frac{\mathbf{z}^H M\mathbf{y}^i}{\mathbf{z}^H M\mathbf{y}^i} - \frac{\mathbf{z}^H A\mathbf{y}^i}{\mathbf{z}^H M\mathbf{y}^i} \\ &= -\frac{\mathbf{z}^H(A - \lambda_1 M)\mathbf{y}^i}{\mathbf{z}^H M\mathbf{y}^i}. \end{aligned}$$

$$= -s^{i+1}\alpha^{i+1}\frac{\mathbf{z}^H(A-\lambda_1 M)\mathbf{u}^{i+1}}{\mathbf{z}^H M\mathbf{y}^i}. \tag{4.36}$$

We use the splitting of $\mathbf{x}^{i+1}$ and the scaling to obtain

$$\mathbf{y}^i = (\mathbf{z}^H M\mathbf{y}^i)\mathbf{x}^{i+1} = (\mathbf{z}^H M\mathbf{y}^i)\alpha^{i+1}(c^{i+1}\mathbf{v}_1^R + s^{i+1}\mathbf{u}^{i+1}).$$

Now if $|\mathbf{z}^H M\mathbf{v}_1^R| > |\mathbf{z}^H M\mathbf{u}^{i+1}|$ and $|c^0| > |s^0|$ then

$$\frac{|\mathbf{z}^H(A-\lambda_1 M)\mathbf{y}^i|}{|\mathbf{z}^H M\mathbf{y}^i|} \leq |s^{i+1}|\,\frac{|\mathbf{z}^H(A-\lambda_1 M)\mathbf{u}^{i+1}|}{|c^0||\mathbf{z}^H M\mathbf{v}_1^R| - |s^0||\mathbf{z}^H M\mathbf{u}^{i+1}|}.$$

Next we define $\widetilde{C}_1 := \max_{\mathbf{u}\neq 0} |\mathbf{z}^H(A-\lambda_1 M)\mathbf{u}|\,(\frac{15}{16}\,|\mathbf{z}^H M\mathbf{v}_1^R| - \frac{1}{4}\,|\mathbf{z}^H M\mathbf{u}|)^{-1}$, then

$$|\lambda_1 - \sigma^{i+1}| \leq |s^{i+1}|\,\widetilde{C}_1.$$

In order to explain why the Wilkinson update is of interest consider $\mathbf{z} = \xi_1\mathbf{v}_1^L + \xi_2\mathbf{w}$ with $|\xi_1|^2 + |\xi_2|^2 = 1$ and $\|\mathbf{w}^H M V_R\| = 1$, then

$$\sigma^{i+1} - \lambda_1 = s^{i+1}\xi_2\alpha^{i+1}\frac{\mathbf{w}^H(A-\lambda_1 M)\mathbf{u}^{i+1}}{\mathbf{z}^H M\mathbf{y}^i}.$$

Obviously reducing $|\xi_2|$ improves the the quality of the Wilkinson update. So if $\mathbf{z}$ is a better approximation of the left eigenvector $\mathbf{v}_1^L$ than $\mathbf{v}_1^R$ is then we expect the Wilkinson update to be a better approximation of the sought eigenvalue than the RQ.

Further we point out that the Wilkinson update reduces in the case of exact linear solves to $\sigma^{i+1} = \sigma^i + 1/(\mathbf{z}^H M\mathbf{y}^i)$, however this is no longer the case for inexact solves, where the update has the form $\sigma^{i+1} = \sigma^i + \mathbf{z}^H(\mathbf{b}^i - \mathbf{res}^i)/(\mathbf{z}^H M\mathbf{y}^i)$.

We now present two variations of Inexact Inverse Iteration using the Wilkinson update, one with fixed tolerance and one with decreasing tolerance.

InvitWf is Algorithm 6 (p. 98) with

$$\mathbf{b}^i = M\mathbf{x}^i \quad \text{and} \quad \tau^i = \tau^0. \tag{4.37}$$

InwitWd is Algorithm 6 (p. 98) with

$$\mathbf{b}^i = M\mathbf{x}^i \quad \text{and} \quad \tau^i = \min\{\tau^0, \widetilde{C}_3\,\|\mathbf{r}^i\|\}. \tag{4.38}$$

Corollary 4.8 *Apply InvitWf, that is Algorithm 6 with (4.37), to the GEP* $A\mathbf{x} = \lambda M\mathbf{x}$ *with* $A, M \in \mathbb{R}^{n\times n}$, *where* M *is spd. Let the initial guess satisfy* $t^0 \leq gap(4\widetilde{C}_1)^{-1}$ *and the residual condition* $\tau^0 \leq 3gap(32\,\|\,V_L^H\,\|\,\widetilde{C}_1)^{-1}$, *then* $t^i \to 0$ *linearly.*

Proof: The proof follows line by line the one of Corollary 4.6 when $\widetilde{\mathcal{R}}$ is replaced by

$\widetilde{C}_1$. □

Corollary 4.9 *Apply InvitWd, that is Algorithm 6 with (4.38), to the GEP* $A\mathbf{x} = \lambda M\mathbf{x}$ *with* $A, M \in \mathbb{R}^{n\times n}$, *where* M *is spd. Let the initial guess satisfy* $t^0 \leq gap(4\widetilde{C}_1)^{-1}$ *and the residual condition* $\tau^0 \leq 3gap(32 \parallel V_L^H \parallel \widetilde{C}_1)^{-1}$, *then* $t^i \to 0$ *and the convergence is locally quadratic.*

Proof: The convergence follows from Corollary 4.9. The proof of the local rate of convergence follows line by line the one of Corollary 4.7 where $\widetilde{\mathcal{R}}$ is replaced by $\widetilde{C}_1$. □

We illustrate the convergence of InvitWf and InvitWd later in Test 4.2, see Section 4.5.

4.3.4 Modified right-hand sides

In this section we extend the method PInvit as introduced in Section 3.6 and based on the approach by Simoncini and Eldén (2002) as well as Scott (1981) to the generalised eigenvalue problem. Until now in this chapter, the methods discussed have used the standard right-hand side $\mathbf{b}^i = M\mathbf{x}^i$. Later in Chapter 6 we show that these methods are not optimal in the sense that GMRES does not benefit from the fact that a good approximation $\mathbf{x}^i$ for the sought solution $\mathbf{y}^{i+1}$ is available. We observed this effect in case of the standard symmetric eigenvalue using preconditioned MINRES in Section 3.6. In case of the symmetric eigenvalue problem we have seen that tailoring the right-hand side $\mathbf{b}^i$ to the linear solver improves the performance of the linear solver to such an extend that the resulting method was most efficient.

Here we provide only the convergence analysis for the method, however we start with a brief motivation for the specific choice for the right-hand side $\mathbf{b}^i = P\mathbf{x}^i$.

Using our abstract notation of Algorithm 5 then solving the linear system

$$(A - \sigma^i M)\mathbf{y}^i = \mathbf{b}^i$$

with preconditioned GMRES then we actually solve the system

$$P_1^{-1}(A - \sigma^i M)P_2^{-1}\mathbf{y}^i = P_1^{-1}\mathbf{b}^i,$$

where P_1 denotes a possible left preconditioner and P_2 a possible right preconditioner. In the remainder we call $P = P_1P_2$ the preconditioner. In Section 3.6.3 we observed that choosing $\mathbf{b}^i$ such that $P_1^{-1}\mathbf{b}^i$ is an approximation of the eigenvector corresponding to the eigenvalue with smallest modulus of $P^{-1}(A - \sigma^i M)P_2^{-1}$ is beneficial for the performance of MINRES. Later in Section 6 we confirm this for the GEP and the use of GMRES as linear solver and also show that $P\mathbf{x}^i$ is an approximation to this eigenvector. Therefore we consider for the remainder of this section the choice $\mathbf{b}^i = P\mathbf{x}^i$.

PInvit is Algorithm 5 (p. 89) with

$$\sigma^i = \varrho^i, \quad \tau^i = \tau^0 \, \|V_L^H M \mathbf{x}^i\|, \quad \mathbf{b}^i = P\mathbf{x}^i, \tag{4.39}$$

We observe that unless $P\mathbf{v}_1^R = \eta M \mathbf{v}_1^R$ for some $\eta \in \mathbb{C}$ we have $\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H P\mathbf{v}_1^R\| \neq 0$ and hence $\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H P\mathbf{x}^i\| \not\to 0$. As a result condition b) in Theorem 4.2 is only satisfied for $\gamma_1 = 0$. Hence the convergence has to be gained by a shift tending towards the desired eigenvalue. We restrict ourselves here to the RQ while the Wilkinson update gives another possible choice for the shift.

We now provide convergence results for this method. The corresponding proof is technical and only presented for completeness.

Corollary 4.10 *Apply PInvit, that is Algorithm 5 with (4.39) to the GEP* $A\mathbf{x} = \lambda M\mathbf{x}$ *with* $A, M \in \mathbb{R}^{n\times n}$ *where* M *is spd. Let the initial guess satisfy*

$$t^0 \;\leq\; \min\{\frac{gap}{4\widetilde{\mathcal{R}}}, \frac{2}{3}\,\|J - \mu I\|^{-1}\},$$

and assume the preconditioner $P = A - \mu M + E$ *is such that* $\|V_L^H E\mathbf{u}\| \leq \frac{1}{3}\,\|V_L^H M\mathbf{u}\|$ *for all* $\mathbf{u}$*, while the stopping condition satisfies* $\tau^0 \leq 3(20\,\|V_L\|)^{-1}$*, then* $t^i \to 0$ *linearly.*

Proof: As $|\lambda_1 - \varrho^i| \leq |s^i|\,\widetilde{\mathcal{R}}$ we set $C_1 = \widetilde{\mathcal{R}}$, and $\beta = 1$ while $\gamma_1 = \gamma_2 = \gamma = 0$. As $t^0 \leq gap(4\widetilde{\mathcal{R}})^{-1}$ condition a) is satisfied. For condition b) we observe that

$$\begin{aligned}
\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H \mathbf{b}^i\| &= \|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H P\mathbf{x}^i\| \\
&= \|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H (A - \mu M + E)\mathbf{x}^i\| \\
&\leq \alpha^i\,|s^i|\,\|V_L^H (A - \mu M)\mathbf{u}^i\| + \|V_L^H E\mathbf{x}^i\| \\
&\leq \alpha^i\,|s^i|\,\|J - \mu I\| + \frac{1}{3}\,\|V_L^H M\mathbf{x}^i\| \\
&\leq \alpha^i\left(\frac{2}{3} + \frac{1}{3}\right) \;\leq\; \alpha^i\frac{145}{144} \;\leq\; \alpha^i\frac{5}{3}\left(\frac{15}{16} - \frac{1}{3}\right) \\
&\leq \alpha^i\frac{5}{3}\left(|c^i| - \frac{1}{3}\,\|V_L^H M\mathbf{x}^i\|\right) \\
&\leq \frac{5}{3}\,|(\mathbf{v}_1^L)^H (A - \mu M + E)\mathbf{x}^i| \\
&= \frac{5}{3}\,|(\mathbf{v}_1^L)^H \mathbf{b}^i|\,.
\end{aligned}$$

Hence with $C_2 = \frac{5}{3}$ condition b) is satisfied. Next for condition c) we obtain

$$\begin{aligned}
\|V_L^H \mathbf{res}^i\| \;\leq\; \|V_L^H\|\,\|\mathbf{res}^i\| \;\leq\; \alpha^i\,\|V_L^H\|\,\tau^0 \;\leq\; \frac{3}{20}\alpha^i \\
\leq\; \frac{1}{4}\alpha^i\left(\frac{15}{16} - \frac{1}{3}\right) \;\leq\; \frac{1}{4}\,|(\mathbf{v}_1^L)^H \mathbf{b}^i|,
\end{aligned}$$

so $C_3 = \frac{1}{4}$. Finally we observe that

$$q = \frac{4C_1 t^0}{3gap}\frac{C_2 + C_3}{1 - C_3} \leq \frac{1}{3}\frac{\frac{5}{3} + \frac{1}{4}}{1 - \frac{1}{4}} \leq \frac{1}{3}\frac{\frac{23}{12}}{\frac{3}{4}} = \frac{23}{27} < 1$$

hence we can use Theorem 4.2, the convergence follows from there. □

In Corollary 4.10 we used the condition $\| V_L^H E\mathbf{u} \| \leq \frac{1}{3} \| V_L^H M\mathbf{u} \|$ which can be rewritten using $\mathbf{u} = V_R\mathbf{z}$ with $\|\mathbf{z}\| = 1$ and $\mathbf{z} \perp \mathbf{e}_1$ as $\|V_L^H E V_R\| \leq \frac{1}{3}$. As the condition $\|V_L^H E V_R\| \leq \frac{1}{3}$ implies $\|E\|$ being small it is not unreasonable to say that Lemma 4.10 is tailored to preconditioners where this error expression is small. In contrast to other conditions as, for example, on the initial guess where for at least close enough t^i the condition holds, this very stringent condition on the preconditioner is active for all t^i. Results tailored for other preconditioners, for example, with $P^{-1}(A - \mu M) = I + E$ where $\|E\|$ is small, have similar conditions, however the proof runs slightly different.

Comparing the conditions of Corollary 4.10 with those in the result for RQIf, Corollary 4.6, we observe that the conditions on the initial guess and the residual tolerance are similar. So the additional condition for the preconditioner restricts the use of this method compared with RQIf.

The possible advantage of the linearly converging PInvit over RQIf is subject of Chapter 6. In Section 4.5, see Test 4.3, we give a few examples on the convergence of PInvit.

4.3.5 Correction Methods

Here we consider two methods using a correction equation to update the eigenvector approximation $\mathbf{x}^i$. The first is the Inverse Correction Method from Rüde and Schmid (1995), the second is the approach from Golub and Ye (2000).

Inverse Correction Method

In Chapter 3 we studied the Inverse Correction Method for the standard symmetric eigenvalue problem. We showed that the Inverse Correction Method is a variation of inexact inverse iteration. Now in the GEP, the Inverse Correction Method is unsurprisingly again a variation of inexact inverse iteration. As the result is proven in the same way as in Chapter 3 we omit the analysis and simply state the convergence result and the algorithm, see Algorithm 7.

For later reference we define the following method.

ICMf is Algortihm 7 with $\sigma^i = \sigma^0$ and $\tau^i = \tau^0$.

Algorithm 7: Inverse Correction Method

Given $\mathbf{x}^0$, and $\varphi(\cdot)$
For $i = 0, 1, 2, \ldots$

- Choose σ^i and τ^i,
- Calculate $\varrho^i := (\mathbf{x}^i)^T A\mathbf{x}^i$ and $\mathbf{r}^i := (A - \varrho^i M)\mathbf{x}^i$,
- Solve $(A - \sigma^i M)\mathbf{z}^i = \mathbf{r}^i$ such that $\|\mathbf{r}^i - (A - \sigma^i M)\mathbf{z}^i\| \leq \tau^i \|\mathbf{r}^i\|$,
- Set $\mathbf{y}^i = \mathbf{x}^i - \mathbf{z}^i$,
- Update $\mathbf{x}^{i+1} = \varphi(\mathbf{y}^i)\mathbf{y}^i$,
- Test for convergence

Lemma 4.11 *Apply ICMf, that is Algorithm 7 with $\sigma^i = \sigma^0$ and $\tau^i = \tau^0$ to the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ with $A, M \in \mathbb{R}^{n\times n}$ where M is spd. Let the shift satisfy $0 < |\lambda_1 - \sigma^0| \leq \frac{1}{4} gap$ and $\sigma^0 \neq \varrho^i$. Further let the residual condition be such that $\tau^i \leq |\varrho^i - \sigma^0|\ (\|V_L\|\ \mathcal{R}^*)^{-1}$ while the initial guess satisfies $t^0 \leq \frac{1}{4}$ then $t^i \to 0$.*

Proof: The proof follows the lines of the proof of Lemma 3.14. □

We expect linear convergence with $t^{i+1} \approx qt^i$ where $q = q_0(1 + C_7\tau^i)$. Here q_0 denotes the rate of convergence as for exact solves.

Later in Section 4.5, see Test 4.4 we illustrate the convergence for ICMf together with the following method.

Golub and Ye

Algorithm 8 published by Golub and Ye (2000) is a natural way of implementing inexact inverse iteration, as the algorithm starts with the previous solution and decreases the residual step by step. All previously discussed methods work with general scaling function φ, so for example $\varphi(\mathbf{y}) = \|\mathbf{y}\|^{-1}$. For Algorithm 8 φ needs to satisfy $\varphi(\psi\mathbf{y}) = \psi\varphi(\mathbf{y})$ for any $\psi \in \mathbb{C}$ with $|\psi| = 1$. While Algorithm 8 is similar to the Inverse Correction Method, the sequences only coincide if the scaling in Algorithm 8 satisfies $\varphi(\mathbf{y}^i) = (\varrho^i - \sigma^0)$, however we can state the following.

Remark 4.12 *The inverse correction method with fixed shift, see Algoritm 7, is a special case of Algorithm 8.*

For later reference we define the following method.

Algorithm 8: Golub and Ye

Given $\mathbf{x}^0 \neq \mathbf{0}$ and σ^0 and $\varphi(\cdot)$,
Set $\mathbf{y}^{-1} = \mathbf{0}$,
For $i = 0, 1, 2, \ldots$

- Calculate $\mathbf{r}_{GY}^i = M\mathbf{x}^i - (A - \sigma^0 M)\mathbf{y}^{i-1}$,
- Choose τ^i,
- Inexact solve $(A - \sigma^0 M)\mathbf{z}^i = \mathbf{r}_{GY}^i$ such that $\|\mathbf{z}^i - (A - \sigma^0 M)\mathbf{d}^i\| \leq \tau^i$,
- Update $\mathbf{y}^i = \mathbf{y}^{i-1} + \mathbf{z}^i$ and $\mathbf{x}^{i+1} = \mathbf{y}^i \varphi(\mathbf{y}^i)$,
- Test for convergence

GY is Algorithm 8 with

$$\tau^i = \min\{\tau^0 \, \|V_L^H M\mathbf{x}^i\|, \widetilde{C}_3 \, \|\mathbf{r}^i\|\}. \tag{4.40}$$

In contrast to ICMf, see Algorithm 7, where we used a *relative* tolerance, we now use for GY, see Algorithm 8, an absolute tolerance. This is done to simplify the convergence result. In practise we use $\tau^i \leq \widetilde{C}_3 \, \|\mathbf{r}^i\|$, see (4.40), which gives us a relative tolerance.

We now relate the update of $\mathbf{y}^i$ to the update in inexact inverse iteration, therefore we write

$$\begin{aligned} \mathbf{y}^i &= \mathbf{y}^{i-1} + \mathbf{z}^i \\ &= \mathbf{y}^{i-1} + (A - \sigma^0 M)^{-1}(M\mathbf{x}^i - (A - \sigma^0 M)\mathbf{y}^i - \mathbf{res}^i) \\ &= (A - \sigma^0 M)^{-1}(M\mathbf{x}^i - \mathbf{res}^i). \end{aligned} \tag{4.41}$$

This enables us to present a convergence result for Algorithm 8.

Corollary 4.13 *Apply GY, that is Algorithm 8 with (4.40), to the GEP* $A\mathbf{x} = \lambda M\mathbf{x}$ *with* $A, M \in \mathbb{R}^{n \times n}$, *where* M *is spd. Let* λ_1 *be simple and gap* > 0, *further let the shift satisfy* $0 < |\lambda_1 - \sigma^0| \leq \frac{1}{4} gap$ *while the initial guess is such that* $t^0 \leq \frac{1}{4}$. *If in (4.40)* $\widetilde{C}_3 \leq (\|V_L\| \, \widetilde{\mathcal{R}})^{-1}$ *then* $t^i \to 0$ *linearly.*

Proof: We use the identity (4.41) for the update and apply Corollary 4.5. □

As with ICMf we expect linear convergence for GY with $t^{i+1} \leq q t^i$, where $q = q_0(1 + \tau^i / |s^i|)$ and q_0 the rate of convergence for exact solves. Golub and Ye (2000) proved convergence for Algorithm 8 with $\varphi(\mathbf{y}^i) = (\mathbf{y}^i)_j$ where the jth component has

the largest modulus, and with $\tau^i = p(\widetilde{q})^i$ for some $p > 0$ and $\widetilde{q} \in (0,1)$. They also provided empirical results that choosing $\widetilde{q} < q$ does not improve the rate of convergence.

Later in Chapter 6 we discuss why solving the residual equation might lead to a competitive algorithm. In Test 4.4, see Section 4.5 we illustrate the convergence for ICMf and GY.

4.4 Methods using the Generalised RQ

In this section we assume that the action of A^H on a vector and that linear solves for

$$(A^H - \sigma M)\mathbf{y} = \mathbf{b} \tag{4.42}$$

are available. Then we can apply inexact inverse iteration to the problem

$$A^H\mathbf{x} = \overline{\lambda} M\mathbf{x}. \tag{4.43}$$

The solution of (4.43) is a left eigenvector of the GEP $A\mathbf{x} = \lambda M\mathbf{x}$, so in our previous notation $\mathbf{x}$ approximates $\mathbf{v}_1^L$. The advantage of calculating approximations to the left and the right eigenvector is due to the fact that the generalised Rayleigh quotient (GRQ) is a better approximation of the sought eigenvalue than the standard Rayleigh quotient is.

First we introduce some additional notation and show that the generalised Rayleigh quotient is a higher order approximation of the sought eigenvalue. Then we state a general algorithm for the two sided approach and consider two practical variations. The first variation we consider is a variation of the RQI, using a fixed tolerance. The other method is PInvitGRQ, which is the approach from Simoncini and Eldén (2002) using the GRQ.

The convergence of these variations follows under certain conditions immediately from Theorem 4.2 as well as the corresponding one sided approach, hence we omit the their proofs.

4.4.1 Notation and basic results

In order to distinguish the approximation of the left eigenvector from the approximation of the right eigenvector we write $\mathbf{x}_L^i$ for the left and $\mathbf{x}_R^i$ for the right eigenvector approximation. Further we use the splitting

$$\begin{aligned} \mathbf{x}_R^i &= \alpha_R^i(c_R^i\mathbf{v}_1^R + s_R^i\mathbf{u}_R^i), \\ \text{and} \quad \mathbf{x}_L^i &= \alpha_L^i(c_L^i\mathbf{v}_1^L + s_L^i\mathbf{u}_L^i), \end{aligned} \tag{4.44}$$

where $\mathbf{u}_R^i$ as in (4.9) and $\mathbf{u}_L^i \in \text{span}\{\mathbf{v}_2^L, \ldots, \mathbf{v}_n^L\}$ and $\|(V_R)^H M \mathbf{u}_L^i\| = 1$. Again the scaling $\alpha_R^i := \|V_L^H M \mathbf{x}_R^i\|$ and $\alpha_L^i := \|V_L^H M \mathbf{x}_L^i\|$ leads to $|c_R^i|^2 + |s_R^i|^2 = |c_L^i|^2 + |s_L^i|^2 = 1$. Now we define the generalised Rayleigh quotient (GRQ) by

$$\varrho_G^i := \frac{(\mathbf{x}_L^i)^H A \mathbf{x}_R^i}{(\mathbf{x}_L^i)^H M \mathbf{x}_R^i}. \tag{4.45}$$

Using the splitting (4.44) and the M-orthogonality of V_L^H and V_R we obtain the following bound for the GRQ

$$\begin{aligned} |\varrho_G^i - \lambda_1| &= \frac{(\mathbf{x}_L^i)^H (A - \lambda_1 M) \mathbf{x}_R^i}{(\mathbf{x}_L^i)^H M \mathbf{x}_R^i} \\ &= \frac{|\alpha_L^i \alpha_R^i| \, |(s_L^i \mathbf{u}_L^i)^H (A - \lambda_1 M) s_R^i \mathbf{u}_R^i|}{|\alpha_L^i \alpha_R^i| \, |\overline{c_L^i} c_R^i (\mathbf{v}_1^L)^H M) \mathbf{v}_1^R + \overline{s_L^i} s_R^i (\mathbf{u}_L^i)^H M) \mathbf{u}_R^i|} \\ &\leq |s_L^i s_R^i| \, \frac{|(\mathbf{u}_L^i)^H (A - \lambda M) \mathbf{u}_R^i|}{|c_L^i c_R^i| - |s_L^i s_R^i|}. \end{aligned}$$

The values $\overline{c_L^i}$ and $\overline{s_L^i}$ denote the complex conjugate values of c_L^i and s_L^i respectively. For the last inequality we used that

$$\begin{aligned} |(\mathbf{u}_L^i)^H M \mathbf{u}_R^i| &= |(\mathbf{u}_L^i)^H M V_R V_L^H M \mathbf{u}_R^i| \\ &\leq \|(\mathbf{u}_L^i)^H M V_R\| \, \|V_L^H M \mathbf{u}_R^i\| = 1. \end{aligned}$$

Similarly we have

$$\begin{aligned} |(\mathbf{u}_L^i)^H (A - \lambda_1 M) \mathbf{u}_R^i| &= |(\mathbf{u}_L^i)^H (A - \lambda_1 M) V_R V_L^H M \mathbf{u}_R^i| \\ &\leq \|(\mathbf{u}_L^i)^H M V_R (J - \lambda_1 I)\| \, \|V_L^H M \mathbf{u}_R^i\| \\ &\leq \|(\mathbf{u}_L^i)^H M V_R\| \, \|J - \lambda_1 I\| \\ &= \|J - \lambda_1 I\| \, . \end{aligned}$$

As the largest singular value of $J - \sigma I$ might be determined by a defective eigenvalue we have according to Lemma 4.1 the bound $\|J - \lambda_1 I\| \leq |\lambda_n - \lambda_1| + 1$. Summarising we obtain for the GRQ

$$|\varrho_G^i - \lambda_1| \leq |s_L^i s_R^i| \, \frac{|\lambda_n - \lambda_1| + 1}{|c_l^i c_R^i| - |s_l^i s_R^i|}. \tag{4.46}$$

If s_L^i and s_R^i simultaneously tend towards zero then the GRQ is a higher order approximation of the sought eigenvalue.

Algorithm 9: Two sided inexact inverse iteration

Given $\mathbf{x}_R^0$ and $\mathbf{x}_L^0$ For $i = 0, 1, 2, \ldots$

- Choose τ_R^i, σ_R^i and $\mathbf{b}_R^i$,
- Solve $(A - \sigma_R^i M)\mathbf{y}_R^i = \mathbf{b}_R^i$ such that $\|\mathbf{b}_R^i - (A - \sigma_R^i M)\mathbf{y}_R^i\| \leq \tau_R^i$,
- Choose τ_L^i, σ_L^i and $\mathbf{b}_L^i$,
- Solve $(A^H - \sigma_L^i M)\mathbf{y}_L^i = \mathbf{b}_L^i$ such that $\|\mathbf{b}_L^i - (A^H - \sigma_L^i M)\mathbf{y}_L^i\| \leq \tau_L^i$,
- Update $\mathbf{x}_R^{i+1}$ and $\mathbf{x}_L^{i+1}$,
- Test for convergence.

4.4.2 Methods

As a first method using the GRQ we consider a two sided version of the Rayleigh quotient iteration with fixed tolerance.

GRQIf is Algorithm 9 with

$$\overline{\sigma_L^i} = \sigma_R^i = \varrho_G^i, \quad \tau_L^i = \tau_R^i = \tau^0, \quad \mathbf{b}_L^i = M\mathbf{x}_L^i, \quad \mathbf{b}_R^i = M\mathbf{x}_R^i,$$
$$\mathbf{x}_L^{i+1} = \mathbf{y}_L^i / \|\mathbf{y}_L^i\| \quad \text{and} \quad \mathbf{x}_R^{i+1} = \mathbf{y}_R^i / \|\mathbf{y}_R^i\| . \tag{4.47}$$

Remark 4.14 *Consider GRQIf, that is Algorithm 9 with (4.47) being applied to the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ with $A, M \in \mathbb{C}^{n\times n}$ where M is spd. Under suitable convergence conditions, similar to those of Corollary 4.6, $t^i \to 0$ and the convergence is locally quadratic.*

Therefore we expect this quadratically converging method to perform as good as the also quadratically converging methods RQId and InvitFd, discussed in Sections 4.3.2 and 4.3.3.

Similarly we could extend the method RQId to use the GRQ, which would lead to cubic convergence. As we have seen in Chapter 2 for the standard symmetric eigenvalue problem, the difference between quadratic and cubic convergence is negligible in practice and so we do not discuss this any further.

Earlier, in Section 4.3.4, we extended the approach from Simoncini and Eldén (2002) to the GEP. There we presented PInvit as a one sided approach, but as we observe

later in Test 4.3 the condition on the preconditioner might be too restrictive. Now we present a two sided version of this method with the aim of overcoming the restriction on the preconditioner by taking advantage of a higher order convergence.

PInvitGRQ is Algorithm 9 with

$$\overline{\sigma_L^i} = \sigma_R^i = \varrho_G^i, \quad \tau_L^i = \tau_R^i = \tau^0, \quad \mathbf{b}_L^i = P_L \mathbf{x}_L^i \quad \mathbf{b}_R^i = P_R \mathbf{x}_R^i,$$
$$\mathbf{x}_L^{i+1} = \mathbf{y}_L^i / \|\mathbf{y}_L^i\|, \quad \text{and} \quad \mathbf{x}_R^{i+1} = \mathbf{y}_R^i / \|\mathbf{y}_R^i\|, \tag{4.48}$$

In the definition of PInvitGRQ P_L denotes the preconditioner for $A^H - \sigma_L^i M$ and P_R the preconditioner for $A - \sigma_R^i M$. Again at this stage we do not need to know if, for example, P_L is a left or a right preconditioner.

Remark 4.15 *Consider PInvitGRQ, that is Algorithm 9 with (4.48) being applied to the GEP $A\mathbf{x} = \lambda M\mathbf{x}$ with $A, M \in \mathbb{C}^{n\times n}$ where M is spd. Under suitable convergence conditions, similar to those of Corollary 4.10, $t^i \to 0$ and the convergence is locally quadratic.*

Now if the initial guess is close enough to the sought eigenvalue the quadratic convergence allows the use of a more moderate residual tolerance and a less accurate preconditioner. However as now two linear solves are needed we expect this method to be less efficient than RQId. We discuss this in more detail later in Chapter 6.

Later in our tests, see Test 4.5 in Section 4.5, we use $P_L = P_R^H$, so that only one preconditioner is needed.

4.5 Tests

4.5.1 Example

Here we consider a small constructed example which gives despite its small size enough insight in the convergence behaviour of the here considered methods.

The matrices A and M are real 62×62 matrices with $M = \text{diag}(1.1, 1.2, 1.3, \ldots, 7.2)$ and $A = VDV^{-1}$ where $V = U + 0.2 * I$ with U a full matrix of uniformly in $(0, 1)$ distributed random variables and $D = diag(D_1, D_2, \ldots, D_6, 1, 2, 3, \ldots, 50)$. Further the matrices D_j are 2×2 real matrices corresponding to the complex eigenvalues of A, $1 \pm 5i$,$1 \pm 1i$, $3 \pm 3i$, $3 \pm 1i$, $5 \pm 5i$ and $5 \pm 1i$. The non-standard construction of V is used to keep a moderate conditioning of the eigenvectors. Figure 4.5.1 shows a plot of the spectrum of the matrix pair A, M, the red asterisks indicate the four eigenvalues of interest.

For our tests we consider four different eigenvalues, two real and two complex ones with one of each well separated and the other an interior eigenvalue. Table 4.1 contains

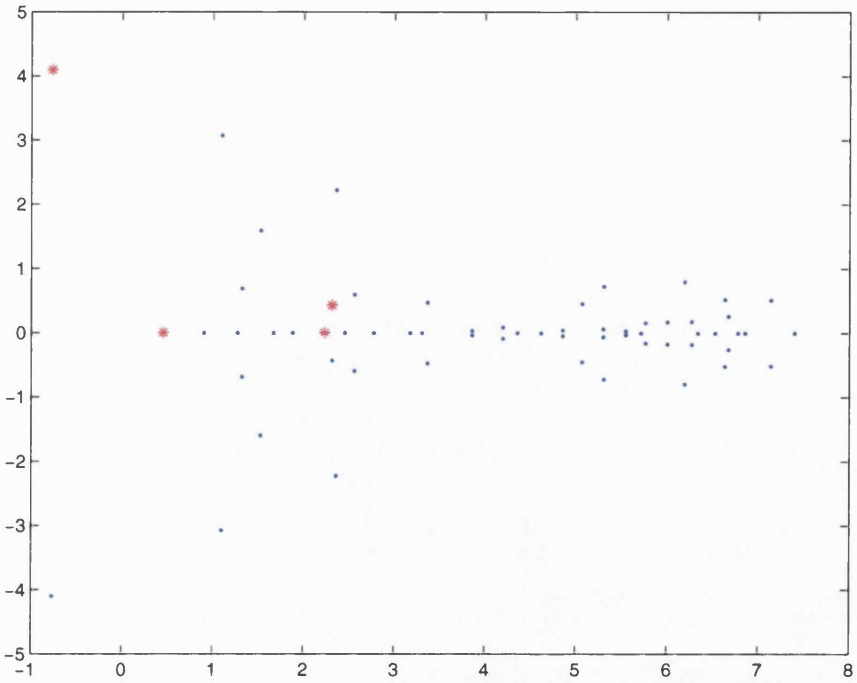


Figure 4-1: Spectrum of the matrix pair (A, M), red stars represent the eigenvalues of interest

	real		complex	
	extreme	interior	extreme	interior
'λ_1'	0.45	2.23	-0.77+4.09i	2.32 + 0.43i
$\frac{\lvert\lambda_1-\sigma^0\rvert}{\lvert\lambda_2-\sigma^0\rvert}$	0.14	0.08	0.06	0.18
gap	0.45	0.22	2.13	0.30

Table 4.1: Eigenvalues of interest and their separation

the eigenvalues of interest and their relative gaps $|\lambda_1 - \sigma^0| \ / \ |\lambda_2 - \sigma^0|$, as well as the corresponding values of *gap*, defined in (4.19).

For the extreme complex eigenvalue which is nicely separated from the remainder we use unpreconditioned GMRES as solver. As unpreconditioned GMRES does not converge sufficiently well for the other eigenvalues we use left preconditioned GMRES. As a preconditioner we use a perturbation of an exact preconditioner. More precisely to precondition the system $(A - \sigma^i M)\mathbf{y}^i = \mathbf{b}^i$ we choose the preconditioner $P^{-1} = (I + E)A^{-1}$. The error matrix is a random matrix with $\|E\| = 0.2$.

For all the tests presented here we used GMRES with two additional stopping conditions. First we check for the targeted eigenvalue residual in GMRES and stop when this target is achieved. Second we use the condition $snprod_k^i \leq 10^{-3}\tau^i$ to detect failure of GMRES before losing all information on the current eigenvector approximation as

	extreme real prec GMRES	interior real prec GMRES	extreme complex GMRES	interior complex prec GMRES
$\widetilde{C}_3$	$0.2/\sigma^0$	$0.2/\sigma^0$	$0.1/\sigma^0$	$0.2/\sigma^0$
i	t^i	t^i	t^i	t^i
0	5.0e-02	5.0e-02	5.0e-02	5.0e-02
1	6.9e-03	1.7e-02	1.1e-02	1.1e-02
2	4.3e-04	7.5e-04	4.8e-04	1.5e-03
3	3.3e-05	4.1e-05	2.5e-05	2.4e-04
4	3.7e-06	2.7e-06	1.5e-06	3.7e-05
5	5.1e-07	2.0e-07	8.8e-08	6.0e-06
6	7.2e-08	1.5e-08	5.4e-09	9.8e-07
7	1.0e-08	1.2e-09	3.3e-10	1.6e-07
8	1.5e-09	9.3e-11	2.1e-11	2.7e-08
9	2.1e-10	7.3e-12	1.3e-12	4.7e-09
10	3.0e-11	6.1e-13	8.2e-14	8.2e-10
11	4.2e-12			1.4e-10
12	6.0e-13			2.5e-11
13	8.8e-14			4.5e-12
14	1.2e-14			7.8e-13

Table 4.2: Convergence history for InvitFd. Tabulated is the tangent t^i against the outer iteration number i for the four eigenvalues of interest, see Table 4.1

is often the case when stopping with the built in tester for stagnation. The variable $snprod_k^i$ is a variable inside the GMRES algorithm and is an estimator for $\|\mathbf{d}_k^i\|$. For more on this see the corresponding discussion on stopping conditions for MINRES in Chapter 3 and for more on GMRES see Chapter 5.

4.5.2 Results

Test 4.1 *We apply InvitFd to calculate the four eigenvalues of interest. The corresponding results are given in Table 4.2.*

We remind ourselves that InvitFd, as defined on page 94, uses a decreasing tolerance $\tau^i \leq \widetilde{C}_3 \|\mathbf{r}^i\|$. As stated in Corollary 4.5 we expect linear convergence for InvitFd.

From Table 4.2 we observe that the convergence of InvitFd is indeed linear and that the rate of convergence is effected by the choice of shift and the choice for τ^i. We explained earlier, see Section 4.3.1, that the rate of convergence should be about $q = q_0(1 + \tau^i / |s^i|)$ where $q_0 = |\lambda_1 - \sigma^i| \ / \ |\lambda_2 - \sigma^0|$ is the rate of convergence for exact solves. For the test we report in Table 4.2 we have $\tau^i / |s^i| \leq 0.5$, therefore we expect no significant slow down of the convergence of InvitFd compared with inverse iteration using exact solves. In fact the empirical convergence rate almost matches with the theoretical convergence rate for exact solves. We also carried out tests with more

	RQIf		RQId	
	interior real		extreme complex	interior complex
	$\tau^0 = 0.1$	$\tau^0 = 0.1$	$\widetilde{C}_3 = 0.1/\sigma^0$	$\widetilde{C}_3 = 0.1/\sigma^0$
i	t^i	t^i	t^i	t^i
0	5.0e-02	5.0e-02	5.0e-02	5.0e-02
1	1.2e-02	5.7e-03	1.7e-01	1.4e-02
2	2.3e-04	6.4e-05	2.7e-02	9.2e-05
3	1.0e-06	1.5e-08	4.3e-04	4.6e-09
4	6.4e-09	6.1e-11	1.7e-07	2.7e-14
5	6.1e-11	3.2e-13	2.1e-14	
6	1.1e-12	6.3e-14	2.0e-14	
7	6.2e-14			

Table 4.3: Convergence history for RQIf and RQId (Test 4.2). Tabulated is the tangent t^i against the outer-iteration number i. For the definition of the three eigenvalues of interest see Table 4.1.

relaxed tolerance conditions but the rate of convergence not necessarily deteriorates. (We do not reproduce those data here.) However this effect is due to the size of our example which results in the actual residual norm being considerably smaller than the tolerance.

Test 4.2 *We apply RQIf and InvitWf to the interior real eigenvalue and RQId and InvitWd to two complex eigenvalues. The results for RQIf and RQId are given in Table 4.3 and those for InvitWf and InvitWd in Table 4.4*

According to the definitions of RQIf, see page 95, and InvitWf, see page 99, we have $\mathbf{b}^i = M\mathbf{x}^i$ and $\tau^i = \tau^0$. The corresponding Corollaries, 4.6 and 4.8 state linear convergence for both methods. In contrast to RQId and InvitWd we expect quadratic convergence, see Corollaries 4.7 and 4.9. The definitions of RQId, see page 95, and InvitWd, see page 99, give $\mathbf{b}^i = M\mathbf{x}^i$ and $\tau^i \leq \min\{\tau^0, \widetilde{C}_3 \, \|\mathbf{r}^i\|\}$.

We expect that the result for RQIf and InvitWf as well as RQId and InvitWd are similar. In all our tests we observed no significant difference with respect to the outer convergence between InvitWf and RQIf and between InvitWd and RQId. Hence we only comment on the data for RQIf and RQId given in Table 4.3. The results for InvitWf and InvitWd are tabulated in Table 4.4.

In Section 4.3.2 we discussed the methods RQIf and RQId and showed that RQIf converges linearly while RQId converges quadratically. The rate of convergence for the linearly converging RQIf depends on the residual constraint for the linear solver, as can be seen in the left two columns of Table 4.3. The same effect is obtained for the other eigenvalues and hence omitted. Also omitted, using a very relaxed tolerance condition might result in no convergence or poor convergence, so, for example, with $\tau^i = 0.5$

	InvitWf		InvitWd	
	interior real		extreme complex	interior complex
	$\tau^0 = 0.1$	$\tau^0 = 0.05$	$\widetilde{C}_3 = 0.5/\sigma^0$	$\widetilde{C}_3 = 0.5/\sigma^0$
i	t^i	t^i	t^i	t^i
0	5.0e-02	5.0e-02	5.0e-02	5.0e-02
1	1.2e-02	5.7e-03	7.7e-03	5.2e-03
2	2.4e-04	6.1e-05	4.9e-05	3.4e-05
3	3.7e-07	6.5e-08	2.4e-09	1.3e-09
4	3.3e-10	3.2e-10	1.1e-14	4.2e-14
5	1.3e-12	1.2e-12		
6	6.2e-14	1.1e-13		

Table 4.4: Convergence history for InvitWf and InvitWd (Test 4.2). Tabulated is the tangent t^i against the outer-iteration number i. For the definition of the two eigenvalues of interest see Table 4.1.

we obtained convergence for the interior real eigenvalue but 16 outer iterations were needed. We observed that for tight enough residual constraints RQIf needs considerably fewer outer-iterations than InvitFd.

The quadratic convergence of RQId is evident in the two right columns of Table 4.3. In both cases the convergence is marginally better than quadratic. However for the extreme complex eigenvalue the estimator for the tangent is larger than the tangent itself and hence the stopping condition was missed by a fraction. This results into an additional outer-iteration. We will return to this problem in Chapter 6, where we consider the efficiency of these methods. Further, in this example the initial guess is so good that the local rate of convergence is observed from the first iteration on.

Test 4.3 *We apply PInvit to the two complex eigenvalues. In Table 4.5 we give results for two test runs each.*

For the definition of PInvit see page 100. In Section 4.3.4 we showed that the convergence of PInvit is linear, see Corollary 4.10, however this result is only applicable for the preconditioned solves, so here for the interior eigenvalue.

Using PInvit on the complex extreme eigenvalue we observed extremely poor convergence which could not be improved by choosing a tighter residual constraint. The rate of convergence for the other eigenvalues was considerably better, but still 15 to 30 iterations were needed to reach the targeted tolerance. While for the previous methods in case of the real eigenvalues linear solves using unpreconditioned GMRES were not feasible, now for PInvit they are feasible.

The convergence of PInvit using unpreconditioned GMRES differs significantly from the convergence of PInvit using preconditioned GMRES. So, for example, using unpreconditioned GMRES and $\tau^0 = 0.1$, PInvit needed only 17 iterations to find a satisfac-

	unpreconditioned		preconditioned	
	extreme complex		interior real	interior complex
τ^0	0.1	0.05	0.1	0.2
	t^i	t^i	t^i	t^i
0	5.0e-02	5.0e-02	5.0e-02	5.0e-02
1	5.3e-02	5.8e-02	1.2e-02	1.4e-02
2	4.2e-02	4.5e-02	2.3e-03	3.1e-03
3	3.4e-02	3.4e-02	1.3e-03	6.6e-04
4	2.6e-02	2.5e-02	6.0e-04	2.6e-04
5	2.1e-02	1.9e-02	2.3e-04	5.4e-05
6	1.7e-02	1.5e-02	1.1e-04	2.1e-05
7	1.3e-02	1.2e-02	4.6e-05	6.3e-06
8	1.0e-02	9.1e-03	2.2e-05	1.1e-06
9	8.4e-03	7.0e-03	8.9e-06	4.8e-07
10	6.6e-03	5.5e-03	4.1e-06	1.2e-07
11	5.2e-03	4.2e-03	1.7e-06	2.3e-08
12	4.1e-03	3.2e-03	7.8e-07	1.1e-08
13	3.3e-03	2.5e-03	3.3e-07	3.4e-09
14	2.6e-03	1.9e-03	1.5e-07	6.8e-10
15	2.0e-03	1.5e-03	6.5e-08	2.5e-10
16	1.6e-03	1.1e-03	2.9e-08	5.0e-11
17	1.3e-03	8.8e-04	1.3e-08	2.6e-11
18	1.0e-03	6.8e-04	5.6e-09	5.0e-12
19	8.1e-04	5.2e-04	2.5e-09	1.8e-12
20	6.4e-04	4.0e-04	1.1e-09	3.6e-13
21	5.1e-04	3.1e-04	4.7e-10	1.9e-13
22	4.0e-04	2.4e-04	2.1e-10	4.9e-14
23	3.2e-04	1.9e-04	9.2e-11	1.8e-14
24	2.5e-04	1.4e-04	4.0e-11	
25	2.0e-04	1.1e-04	1.8e-11	
26	1.6e-04	8.5e-05	7.9e-12	
27	1.3e-04	6.6e-05	3.4e-12	
28	1.0e-04	5.1e-05	1.5e-12	
29	8.0e-05	3.9e-05	6.8e-13	
30	6.3e-05	3.0e-05	3.1e-13	

Table 4.5: Convergence history for PInvit (Test 4.3). Tabulated is the tangent t^i against the outer-iteration number i. For the definition of the two eigenvalues of interest see Table 4.1.

tory approximation of the sought eigenvector. In contrast PInvit using preconditioned GMRES needed 30 iterations. Further for the complex extreme eigenvalue PInvit using preconditioned GMRES failed even for exact solves. This is not surprising as the preconditioner does not satisfy the condition in Corollary 4.10 and the preconditioner is not tailored to this complex extreme eigenvalue.

We also observe the somehow surprising effect that the convergence for the complex interior eigenvalue is considerably better than for the real interior eigenvalue. So far we have no explanation for this effect.

Further we point out that in contrast to the standard symmetric eigenvalue problem the convergence of PInvit is not independent of the residual constraint τ^0. While for the standard symmetric case $\tau^0 = 0.8$ lead to convergence, here we need smaller values for τ^0 and for large problems considerably smaller values might be necessary.

Test 4.4 *We apply ICMf and GY to calculate the extreme real and the complex interior eigenvalue, the corresponding results are given in Table 4.6.*

In Section 4.3.5 we discussed ICMf and GY and stated their linear convergence, see Corollaries 4.11 and 4.13. From the definition of the methods we see that for ICMf $\tau^i = \tau^0$, see page 102, and for GY $\tau^i = \widetilde{C_3}\,\|\mathbf{r}^i\|$, see page 104.

Comparing Table 4.6, containing the results for ICMF and GY, with Table 4.2, containing the results for InvitFd, we observe the same rate of convergence and that there is no significant difference between ICMf and GY. In contrast to the methods using the standard right-hand side, $\mathbf{b}^i = M\mathbf{x}^i$, that are InvitFd, RQIf, RQId, InvitWf and InvitWd, unpreconditioned GMRES converges also for the real eigenvalue problem and the complex interior one.

Test 4.5 *We apply GRQIf and PInvitGRQ to the extreme real and the interior complex eigenvalue, the results are given in Table 4.7.*

In section 4.4.2 we discussed the quadratically converging methods GRQIf and PInvitGRQ see Remarks 4.14 and 4.15. In contrast to the one sided RQIf and RQId which are effected by the conditioning of the sought eigenvalue, GRQIf and PInvitGR are robust even for poorly conditioned eigenvalues.

In all tests carried out with GRQIf we used fewer than 5 iterations to obtain a satisfactory approximation of the sought eigenpair. Comparing the result for GRQIf, see left columns of Table 4.7, and RQId, see right columns in Table 4.3, we observe no significant difference for the here considered example.

In tests with PInvitGRQ we needed occasionally 6 or 7 iterations due to poorer initial convergence otherwise the convergence is excellent. Specially we did not encounter any of the convergence problems as we did using PInvit, see Test 4.3. The difficulty with the initial iteration is apparent from the data given in the right columns of Table 4.7.

4.6 Conclusions

Based on the Jordan decomposition we analysed the convergence of inexact inverse iteration applied to the GEP independent of any specific linear solver. We extended the definition of the tangent of the error angle made in Chapter 2 to the unsymmetric generalised eigenvalue problem and called it the generalised tangent. Based on the Jordan decomposition we established a one-step bound on the generalised tangent for the next iterate. The key result in this chapter, Theorem 4.2, is based on the one-step bound (4.28). Both, Theorem 4.2 and the one step bound (4.28) are key to the efficiency analysis presented later in Chapter 6. We point out that Theorem 4.2 is general in the sense that it is valid with any (iterative) linear solver satisfying the residual constraints. Further, Theorem 4.2 is also general in the sense that it is applicable to many if not all variations of inexact inverse iteration. Corollary 4.3 applies Theorem 4.2 to the important and more widely known variations of inexact inverse iteration using the standard right-hand side, such as Invit, RQIf and RQId. Based on Theorem 4.2, Corollary 4.3 and Remark 4.4 we considered a few variations of inexact inverse iteration for which we concluded convergence, super linear convergence and the type of superlinear convergence. Among the methods studied were methods using a fixed shift, or a shift equaling the RQ. Further, in Section 4.3.4 we extended the approach from Simoncini and Eldén (2002) to the generalised nonsymmetric eigenvalue problem. Finally we discussed a few approaches based on the generalised Rayleigh quotient.

Using a small constructed example we illustrated the convergence of these methods. With respect to the outer convergence, methods with superlinear convergence that are RQId, InvitWd, GRQIf and PInvitGRQ outperformed the other methods. However this ignores the performance of the inner-method and thus does not allow a fair judgement of which method should be recommended for practical use.

The tests carried out were mainly designed to show the difference between the methods rather than a test for robustness. So that in order to understand practical limitations of the methods more tests specially on large problems might be useful.

Later in Chapter 6 we consider GMRES as linear solver and question which of the methods studied here is efficient.

	ICMf		GY	
	extreme real	interior real	extreme real	interior real
	$\tau^0 = 0.1$	$\tau^0 = 0.1$	$\widetilde{C}_3 = 0.1/\sigma^0$	$\widetilde{C}_3 = 0.1/\sigma^0$
i	t^i	t^i	t^i	t^i
0	5.0e-02	5.0e-02	5.0e-02	5.0e-02
1	2.1e-03	1.2e-02	1.9e-03	5.7e-03
2	3.2e-04	7.0e-04	3.7e-04	5.0e-04
3	3.7e-05	6.2e-05	4.2e-05	2.2e-05
4	5.5e-06	1.0e-05	5.8e-06	1.8e-06
5	7.6e-07	7.2e-07	8.3e-07	2.1e-07
6	1.1e-07	5.6e-08	1.2e-07	9.6e-09
7	1.6e-08	4.3e-09	1.7e-08	2.8e-09
8	2.2e-09	3.7e-10	2.4e-09	2.5e-10
9	3.1e-10	2.7e-11	3.4e-10	1.8e-11
10	4.4e-11	1.8e-12	4.8e-11	1.3e-12
11	6.3e-12	1.9e-13	6.8e-12	1.5e-13
12	8.9e-13	5.6e-14	9.7e-13	4.4e-14
13	1.3e-13		1.4e-13	
14	1.8e-14		1.9e-14	
15	3.6e-15		3.8e-15	

Table 4.6: Convergence history for ICMf and GY (Test 4.4). Tabulated is the tangent t^i against the outer-iteration number i. For the definition of the two eigenvalues of interest see Table 4.1.

	GRQIf		PInvitGRQ	
	extreme real	interior complex	extreme real	interior complex
τ^0	0.2	0.2	0.1	0.2
i	t^i	t^i	t^i	t^i
0	5.0e-02	5.0e-02	5.0e-02	5.0e-02
1	4.3e-03	4.2e-02	9.2e-01	2.4e-01
2	1.0e-06	2.9e-03	2.5e-01	1.0e-01
3	7.5e-14	9.5e-07	3.2e-03	2.3e-03
4		2.9e-13	1.1e-06	7.2e-07
5			2.1e-14	2.3e-14

Table 4.7: Convergence history for GRQIf and PInvitGRQ (Test 4.5). Tabulated is the tangent t^i against the outer-iteration number i. For the definition of the two eigenvalues of interest see Table 4.1.

Chapter 5

GMRES

In the previous chapter we have seen that in inexact inverse iteration systems of the form $(A - \sigma M)\mathbf{y} = \mathbf{b}$ arise. As we go on to analyse the efficiency of inexact inverse iteration using GMRES as a linear solver later in Chapter 6 we need to know how GMRES and preconditioned GMRES performs when applied to systems of the above form. The main result for this will be Corollary 5.11.

As the shift σ might be complex we consider first general complex systems $B\mathbf{y} = \mathbf{b}$ before later in Section 5.2.3 returning to the special case where $B = A - \sigma M$, with A and M real.

Introduced by Saad and Schultz (1986), GMRES is an iterative Galerkin-Krylov technique for solving linear systems. Originally published for real unsymmetric systems it can handle complex valued systems $B\mathbf{y} = \mathbf{b}$, where $B \in \mathbb{C}^{n\times n}$ and $\mathbf{y}, \mathbf{b} \in \mathbb{C}^n$. Earlier in Section 3.2 we discussed MINRES, which is also a Galerkin-Krylov technique, and many of the remarks made there apply also for GMRES, as MINRES can be viewed as a special implementation of GMRES for symmetric systems. In order to avoid any confusion we will repeat those here, or state explicitly that they apply.

The basic idea in GMRES is to find in each iteration the vector $\mathbf{y}_k$ solving

$$\|\mathbf{b} - B\mathbf{y}_k\|_2 \;=\; \min_{\mathbf{y}\in\mathcal{K}_k} \|\mathbf{b} - B\mathbf{y}\|_2, \tag{5.1}$$

where $\mathcal{K}_k$ denotes the Krylov-subspace $\mathcal{K}_k := \text{span}\{\mathbf{b}, B\mathbf{b}, \ldots, B^{k-1}\mathbf{b}\}$. Using the set

$$\Pi_k^1 \;:=\; \{f | f \text{ polynomial with degree}(f) \leq k \text{ and } f(0) = 1\} \tag{5.2}$$

we can write the minimisation problem (5.1) as

$$\|\mathbf{b} - B\mathbf{y}_k\|_2 \;=\; \min_{f\in\Pi_k^1} \|f(B)\mathbf{b}\|_2\,. \tag{5.3}$$

This polynomial formulation is the standard way to study the convergence of GMRES

using constrained minimal polynomials, Π_k^1, over the eigenvalues of B. However, Greenbaum et al. (1996) showed that the eigenvalues of the system matrix B on their own are not enough to bound the convergence of GMRES. For a given set of eigenvalues and a given convergence curve they construct a matrix and corresponding right-hand side, such that $\|\mathbf{d}_k\|_2 := \|\mathbf{b} - B\mathbf{y}_k\|_2$ behaves according to the prescribed convergence curve, hence any non-increasing convergence curve is possible. However, their result uses the departure from normality of the system matrix B, which affects the conditioning of the eigenvectors. Given the eigenvalues and the eigenvectors of B the convergence can be bounded using a polynomial on the eigenvalues. We follow this standard approach here. That the resulting bound describes the convergence behaviour has been demonstrated by Embree (1999) and Liesen (2000).

Before we present the convergence analysis in Section 5.2 we discuss in Section 5.1 the minimisation problem

$$\eta_D^k := \min_{f \in \Pi_k^1} \max_{z \in D} |f(z)|, \tag{5.4}$$

where D is a non-empty and compact subset of $\mathbb{C}$. Finally in Section 5.3 we briefly discuss further literature on GMRES and variations of the algorithm.

5.1 Constrained Minimal Polynomials

The constrained minimisation problem (5.4) for complex domains has been discussed for example in Manteuffel (1977); Fischer and Freund (1990, 1991); Chatelin (1993); Fischer and Peherstorfer (2001). Here we only summarise those results which are of interest to our application. First, in Section 5.1.1, we discuss some cases where $\eta_D^k \not\to 0$ for $k \to \infty$. Then in Section 5.1.2 we use a result from Fischer and Freund (1990) to bound η_D^k, where D is an ellipse. Based on a result in Chatelin (1993), we give in Section 5.1.3 a bound on η_D^k for the case where D is a disk. To obtain bounds for more complicated disconnected domains we use polynomial maps. Finally, in Section 5.1.5 we discuss arbitrary domains D for which $\eta_D^k \to 0$ and show that the convergence is at least linear in k.

5.1.1 Domains with holes

We start with the obvious observation that for the case where $0 \in D$ the minimising solution is $p(z) \equiv 1$ and so $\eta_D^k = 1$ for all k. Another difficulty arises if $\mathbb{C}\backslash D$ is disconnected, and 0 and ∞ are in disconnected subsets. To illustrate the difficulty consider the unit circle $D = \{z| \ |z| = 1\}$ then $\eta_D^k = 1$ for all k. This follows from the fact that polynomials are holomorphic and holomorphic functions can not have a maximum in the interior of a set, see Ablowitz and Fokas (1997, Theorem 2.6.6, p. 97).

5.1.2 Ellipse

Consider an ellipse with foci f_1, and f_2 and radius $R \in \mathbb{R}^+$,

$$\mathcal{E}_{R,f_1,f_2} := \{z| \ |f_1 - z| + |f_2 - z| \leq 2R\}, \quad f_1, f_2 \in \mathbb{C} \tag{5.5}$$

with $0 \notin \mathcal{E}_{R,f_1,f_2}$, and $R > \frac{1}{2}|f_2 - f_1|$. Later we consider such general elliptical domains, however standard results are written in terms of a standard ellipse $\mathcal{E}_r$ with real foci at 1 and --1, i.e.

$$\mathcal{E}_r := \mathcal{E}_{r,1,-1} = \{z| \ |1 - z| + |-1 - z| \leq 2r\}. \tag{5.6}$$

The transformation

$$g(z) := \frac{f_1 + f_2 - 2z}{f_1 - f_2} \tag{5.7}$$

maps $\mathcal{E}_{R,f_1,f_2}$ to $\mathcal{E}_r$ with $r = 2R/|f_2 - f_1|$.

Lemma 5.1 *Consider the ellipse $\mathcal{E}_r$ defined in (5.6) and $z_0 \in \mathbb{C}\backslash\mathcal{E}_r$. Further let $\beta \in \mathbb{R}^+$ such that $\mathbf{z}_0 \in \partial\mathcal{E}_\beta$ then*

$$\min_{\substack{f \in \Pi_k \\ f(z_0)=1}} \max_{z \in \mathcal{E}_r} |f(z)| \leq \frac{T_k(r)}{T_k(\beta)}.$$

Proof: This result is proven in Fischer and Freund (1990, Theorem 2 and equation (1)). As their notation differs we point out that $\mathbf{z}_0 \in \mathbb{C}\backslash\mathcal{E}_r$ implies $\beta > r$. Further their right-hand side quotient is gained by using the Joukowsky map, i.e. $T_k(r) = \frac{1}{2}(\tilde{r}^k + 1/\tilde{r}^k)$ for $2r = \tilde{r} + 1/\tilde{r}$. □

A similar result can be found in Chatelin (1993, Lemma 7.3.1), for general real foci $f_1, f_2, z_0 \in \mathbb{R}$. Further Fischer and Freund (1990) show that the bound is attained for β large enough, i.e. $\beta \gg r$. However, in our application the spectrum of the possibly preconditioned matrix might have a much less favourable relative separation of the origin and hence $\beta \approx r$. For this case the bound is not sharp as pointed out by Fischer and Freund (1991).

Now we present a Corollary to Lemma 5.1 giving a more practical bound by bounding the Chebyshev polynomials.

Corollary 5.2 *Consider the minimisation problem (5.4) on the ellipse $D = \mathcal{E}_{R,f_1,f_2}$, see (5.5). If $|f_1 - f_2| < 2R < |f_1| + |f_2|$ then $0 \notin D$ and*

$$\eta_D^k \leq \frac{T_k\left(\frac{2R}{|f_2 - f_1|}\right)}{T_k\left(\frac{|f_2| + |f_1|}{|f_2 - f_1|}\right)} \tag{5.8}$$

$$\leq 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\,\frac{\sqrt{\kappa_1}+1}{\sqrt{\kappa_1}-1}\right)^k, \tag{5.9}$$

where

$$\kappa_1 := \frac{2R+|f_2-f_1|}{2R-|f_2-f_1|} \quad \textit{and} \quad \kappa_2 := \frac{|f_1|+|f_2|+|f_2-f_1|}{|f_1|+|f_2|-|f_2-f_1|}.$$

Proof: We start for (5.8) by using $g(z)$ as defined in (5.7) and observe

$$\begin{aligned}
\eta_D^k &= \min_{p\in\Pi_k^1} \max_{z\in D} |p(z)| \\
&= \min_{\substack{p\in\Pi_k \\ p(g(0))=1}} \max_{z\in D} |p(g(z))| \\
&= \min_{\substack{p\in\Pi_k \\ p(z_0)=1}} \max_{z\in\mathcal{E}_r} |p(z)| \\
&\leq \frac{T_k(r)}{T_k(\beta)},
\end{aligned}$$

where $z_0 = g(0) = \frac{f_1+f_2}{f_1-f_2}$ and β such that $z_0 \in \partial\mathcal{E}_\beta$, while $r = 2R/\,|f_1-f_2|$. To obtain the explicit value of β we use that $z_0 \in \partial\mathcal{E}_\beta$, hence $|1-z_0| + |-1-z_0| = 2\beta$ which is the same as $\beta = (|f_1|+|f_2|)/\,|f_1-f_2|$.

In order to obtain the bound, (5.9), we use the fact that

$$\left(\frac{\sqrt{\frac{x+y}{x-y}}+1}{\sqrt{\frac{x+y}{x-y}}-1}\right)^k > T_k\left(\frac{x}{y}\right) > \frac{1}{2}\left(\frac{\sqrt{\frac{x+y}{x-y}}+1}{\sqrt{\frac{x+y}{x-y}}-1}\right)^k,$$

see (A.12). □

So far we did not show that the term on the right-hand side in (5.9) tends to zero. To achieve this we only need to show that the term in the brackets is in $(0,1)$. As $0 \notin D$ we obtain $|f_1|+|f_2| \geq |f_1+f_2| > 2R$ and hence $1 < \kappa_2 < \kappa_1$. Using the monotonicity of $\sqrt{\ }$ we obtain

$$\begin{aligned}
\sqrt{\kappa_2} &< \sqrt{\kappa_1} \\
\Leftrightarrow \sqrt{\kappa_2}-\sqrt{\kappa_1} &< \sqrt{\kappa_1}-\sqrt{\kappa_2} \\
\Leftrightarrow \sqrt{\kappa_2}\sqrt{\kappa_1}+\sqrt{\kappa_2}-\sqrt{\kappa_1}-1 &< \sqrt{\kappa_2}\sqrt{\kappa_1}+\sqrt{\kappa_1}-\sqrt{\kappa_2}-1 \\
\Leftrightarrow (\sqrt{\kappa_2}-1)(\sqrt{\kappa_1}+1) &< (\sqrt{\kappa_2}+1)(\sqrt{\kappa_1}-1) \\
\Leftrightarrow \frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\frac{\sqrt{\kappa_1}+1}{\sqrt{\kappa_1}-1} &< 1,
\end{aligned}$$

therefore Corollary 5.2 proves that $\eta_D^k \to 0$ with $k \to \infty$.

For more on Chebyshev polynomials see Appendix A.

5.1.3 Disk

Here we consider disks of the form

$$C_{r,c} := \{z| \ |z-c| \leq r\}, \quad 0 < r < |c|, \ c \in \mathbb{C}. \tag{5.10}$$

We now extend a result from Chatelin (1993, Theorem 7.3.4), for disks with real centers $c \in \mathbb{R}$, to complex valued centers $c \in \mathbb{C}$.

Lemma 5.3 *Let* $D = \{z| \ |z-c| \leq r\}$ *with* $0 < r < |c|$ *and* $0 \notin D$ *then* $\eta_D^k = (r/\ |c|)^k$.
Proof: We use Chatelin (1993, Corollary 7.3.5), which states for $z_0, \widetilde{c}, \widetilde{r} \in \mathbb{R}$ with $z_0 > \widetilde{c} + \widetilde{r}$ that

$$\min_{\substack{f \in \Pi_k \\ f(z_0)=1}} \max_{|z-\widetilde{c}| \leq \widetilde{r}} |f(z)| = \left(\frac{\widetilde{r}}{|z_0 - \widetilde{c}|}\right)^k.$$

Define $g(z) := 1 - z/c$, then $g(D) = \{z| \ |z| \leq \widetilde{r}\}$ with $\widetilde{r} = r/\ |c|$ and $\widetilde{c} = 0$, further $z_0 = g(0) = 1$ and therefore

$$\eta_D^k = \min_{\substack{f \in \Pi_k \\ f(g(0))=1}} \max_{z \in g(D)} |f(z)| = \left(\frac{\widetilde{r}}{|1-0|}\right)^k = \left(\frac{r}{|c|}\right)^k.$$

□

5.1.4 Polynomial maps

So far we have only bounds for η_D^k if D is either a disk or an ellipse. Here we use a simple observation to extend the results for disks and ellipses to domains which map into a disk or an ellipse using a polynomial map. The observation we use is that for any $\varphi \in \Pi_m$ and $f \in \Pi_k$ the composition $f \circ \varphi \in \Pi_{km}$, hence

$$\min_{\substack{f \in \Pi_{km} \\ f(z_0)=1}} \max_{z \in D} |f(z)| \leq \min_{\substack{g \in \Pi_k \\ g(\varphi(z_0))=1}} \max_{z \in D} |g(\varphi(z))|. \tag{5.11}$$

$$= \min_{\substack{g \in \Pi_k \\ g(\varphi(z_0))=1}} \max_{z \in \varphi(D)} |g(z)|. \tag{5.12}$$

The idea is to use a polynomial map φ, i.e. φ is polynomial, such that $\varphi^{-1(D)}$ is a suitable disk or ellipse while D might be a disconnected set.

Above idea is more carefully studied in Fischer and Peherstorfer (2001), specifically with respect to the quality of the inequality (5.11). For any given set D and any $k \in \mathbb{N}$

let $T_k^D(z)$ denote the solution of the Chebyshev approximation problem

$$\max_{z\in D} |T_k^D(z)| \;=\; \min_{f\in\Pi} \max_{z\in D} |z^n - f(z)| \,.$$

Then Fischer and Peherstorfer (2001, Corollary 2.2) state

$$T_{km}^{\varphi^{-1}(D)}(z) \;=\; a^{-k}\, T_k^D(\varphi(z)).$$

for $\varphi \in \Pi_k \backslash \Pi_{k-1}$ with m simple zeros and leading coefficient a.

We now apply above idea to Corollary 5.2 and Lemma 5.3 to obtain following Lemma, for which we did not find a suitable reference.

Lemma 5.4 *Consider the minimisation problem (5.4). Let $\varphi \in \Pi_m$ with $\varphi(0) = 0$. If $D \subset \mathbb{C}$ is such that $0 \notin \varphi(D)$ and*

1. *that $\varphi(D) \subset \mathcal{E}_{R,f_1,f_2}$, see (5.5), then $\eta_D^k \le p\, q^k$ where*

$$q \;=\; \left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\,\frac{\sqrt{\kappa_1}+1}{\sqrt{\kappa_1}-1}\right)^{1/m} \quad \text{and} \quad p \;=\; 2/q^m, \tag{5.13}$$

   where κ_1 and κ_2 as defined in Corollary 5.2.

2. *that $\varphi(D) \subset \mathcal{C}_{r,c}$, see (5.10), then $\eta_D^k \le p\, q^k$ where*

$$q \;=\; (r/\,|c|)^{1/m} \quad \text{and} \quad p \;=|c|\,/r. \tag{5.14}$$

Proof: For the ellipse, Corollary 5.2, as for the disk, Lemma 5.3, we have the bound

$$\eta_{\varphi(D)}^j \;\le \widetilde{p}\,\widetilde{q}^j,$$

$$\eta_{\varphi(D)}^{j'} \le \widetilde{p}\,\widetilde{q}^{j},$$

where for example $\widetilde{p} = 1$ and $\widetilde{q} = r/\,|c|$ for the circle. Using (5.11) we have $\eta_D^{jm} \le \eta_{\varphi(D)}^j \le \widetilde{p}\,\widetilde{q}^j$. By setting $q = (\widetilde{q})^{1/m}$ and $p = \widetilde{p}/\widetilde{q}$ we obtain

$$\eta_D^k \;\le\; \eta_D^{[k/m]m} \;\le\; \eta_{\varphi(D)}^{[k/m]} \;\le\; \widetilde{p}\,\widetilde{q}^{[k/m]} \;\le\; p\, q^{[k/m]m+m} \;\le\; p\, q^k.$$

5.1.5 General Domains

Earlier in Section 5.1.1 we gave an example of a domain D where the minimisation problem (5.4) does not change with k, i.e. $\eta_D^k = 1$ for all $k \in \mathbb{N}$. Here we consider any

set $D \subset \mathbb{C}$ with $\eta_D^k \to 0$. If $\varphi(z) \in \Pi^1_{k^*}$ then $(\varphi(z))^m \in \Pi^1_{k^*m}$ and therefore

$$\begin{aligned} \eta_D^{k^*m} &= \min_{f \in \Pi^1_{k^*m}} \max_{z \in D} |f(z)| \\ &\leq \min_{f \in \Pi^1_{k^*}} \max_{z \in D} |(f(z))^m| = (\eta_D^{k^*m})^m. \end{aligned} \tag{5.15}$$

We use this observation to obtain the following result.

Lemma 5.5 *Consider the minimisation problem (5.4), where $D \subset \mathbb{C}$ is such that $\exists k^*$ with $\eta_D^{k^*} < 1$, then $\eta_D^k \leq pq^k$ where $q = (\eta_D^{k^*})^{1/k^*}$ and $p = 1/\eta_D^{k^*}$.*

Proof: We use (5.15) together with the technique used in the proof of Lemma 5.4,

$$\eta_D^k \leq \eta_D^{[k/m]m} \leq (\eta_D^m)^{[k/m]} \leq (q^m)^{[k/m]} \leq p\, q^{m[k/m]+m} \leq p\, q^k. \tag{5.16}$$

□

5.2 Convergence

5.2.1 Standard Anaysis

In this section we analyse the convergence of GMRES applied to complex valued linear systems $B\mathbf{y} = \mathbf{b}$. We assume that we can choose the set $D \subset \mathbb{C}$, containing all but a few eigenvalues of B, such that $\eta_D^k \to 0$. Eigenvalues which might be treated separately include eigenvalues close to the origin, defective eigenvalues and eigenvalues where eigenvectors are badly conditioned. As in Chapter 3 for the real symmetric eigenvalue problem, our main interest is to treat eigenvalues separately if thereby D can be chosen such that q as in Lemma 5.5 can be reduced significantly. This common technique can, for example, be found in Hackbusch (1994, Section 7.3.6).

In order to state the key result we introduce the residual after the kth iteration of GMRES

$$\mathbf{d}_k := \mathbf{b} - B\mathbf{y}_k. \tag{5.17}$$

Further for a given set $\Gamma \subset \mathbb{N}_n$ define

$$Q_\Gamma := \operatorname{diag}(\delta_1(\Gamma), \ldots, \delta_n(\Gamma)), \tag{5.18}$$

where $\delta_j(\Gamma) = 0$ if $j \in \Gamma$ and $\delta_j(\Gamma) = 1$ otherwise. Let $\mu_1, \ldots, \mu_n$ denote the eigenvalues of B then we use $D_\Gamma \supset \{\mu_j | j \notin \Gamma\}$ and will assume that $\eta_{D_\Gamma}^k \to 0$.

Lemma 5.6 *Consider GMRES applied to $B\mathbf{y} = \mathbf{b}$ where $B \in \mathbb{C}^{n \times n}$ and $\mathbf{b} \in \mathbb{C}^n$. Let B be non-singular and have the eigenvalue decomposition $B = WJW^{-1}$ with eigenvalues*

$\mu_1, \ldots, \mu_n$. Let $\Gamma \subset \mathbb{N}_n$ with $1 \in \Gamma$ and $\{j | \mu_j$ defective$\} \subset \Gamma$ be such that there exist D_Γ with $\eta_{D_\Gamma}^k \to 0$ for $k \to \infty$. Then there exists $p > 0$ and $q \in (0,1)$ as in Lemma 5.5 that for all $k \geq |\Gamma|$

$$\|\mathbf{d}_k\|_2 \;:=\; \|WQ_\Gamma\|_2 \, p_\Gamma \, {q_\Gamma}^{k-|\Gamma|} \, |\mu_1|^{-1} \|Q_\Gamma W^{-1}\mathbf{b}\|_2, \tag{5.19}$$

where $q_\Gamma = q$ and

$$\widetilde{p_\Gamma} \;:=\; p \max_{\xi \in D_\Gamma} \left(|\mu_1 - \xi| \prod_{j \in \Gamma \setminus \{1\}} \frac{|\mu_j - \xi|}{|\mu_j|} \right).$$

Proof: We define the auxiliary polynomial $g(\xi) := \prod_{j\in\Gamma}(\mu_j - \xi)/\mu_j$ and use that JQ_Γ is diagonal, then $g(J) = g(JQ_\Gamma) = g(J)Q_\Gamma = Q_\Gamma g(J)$. Further as JQ_Γ is diagonal $f(J)Q_\Gamma$ is diagonal and $Q_\Gamma f(J) = f(JQ_\Gamma) = f(J)Q_\Gamma$. Hence

$$\begin{aligned}
\|\mathbf{d}_k\|_2 &= \min_{f\in\Pi_k^1} \|f(B)\mathbf{b}\|_2 \\
&= \min_{f\in\Pi_k^1} \|Wf(J)W^{-1}\mathbf{b}\|_2 \\
&\leq \min_{f\in\Pi_{k-|\Gamma|}^1} \|Wf(J)g(J)W^{-1}\mathbf{b}\|_2 \\
&= \min_{f\in\Pi_{k-|\Gamma|}^1} \|WQ_\Gamma f(J)g(J)Q_\Gamma W^{-1}\mathbf{b}\|_2 \\
&\leq \min_{f\in\Pi_{k-|\Gamma|}^1} \|WQ_\Gamma\|_2 \|f(JQ_\Gamma)\|_2 \|g(JQ_\Gamma)\|_2 \|Q_\Gamma W^{-1}\mathbf{b}\|_2 \\
&\leq \|WQ_\Gamma\|_2 \min_{f\in\Pi_{k-|\Gamma|}^1} \max_{j\in\Gamma^C} |f(\mu_j)| \max_{j\in\Gamma^C} |g(\mu_j)| \|Q_\Gamma W^{-1}\mathbf{b}\|_2 \\
&\leq \|WQ_\Gamma\|_2 \min_{f\in\Pi_{k-|\Gamma|}^1} \max_{\xi\in D_\Gamma} |f(\xi)| \max_{\xi\in D_\Gamma} |g(\xi)| \|Q_\Gamma W^{-1}\mathbf{b}\|_2 \\
&\leq \|WQ_\Gamma\|_2 \, p \, {q_\Gamma}^{k-|\Gamma|} \max_{\xi\in D_\Gamma} \left| (\mu_1 - \xi) \prod_{j\in\Gamma\setminus\{1\}} \frac{\mu_j - \xi}{\mu_j} \right| |\mu_1|^{-1} \|Q_\Gamma W^{-1}\mathbf{b}\|_2 \\
&\leq \|WQ_\Gamma\|_2 \, \widetilde{p_\Gamma} \, {q_\Gamma}^{k-|\Gamma|} \, |\mu_1|^{-1} \|Q_\Gamma W^{-1}\mathbf{b}\|_2 \, .
\end{aligned}$$

□

The assumption that the indices of all defective eigenvalues are in Γ makes this result less suitable for matrices W with many defective eigenvalues. The condition $\eta_D^k \to 0$ for $k \to \infty$ can be satisfied by taking suitable eigenvalue indices into Γ.

In the standard literature like Saad and Schultz (1986) and Saad (1996) Lemma 5.6 is presented for $\Gamma = \{\}$, hence (5.19) has the form

$$\|\mathbf{d}_k\|_2 \;\leq\; \kappa_2(W) \min_{f\in\Pi_k^1} \max_{j\in\mathbb{N}_n} |f(\mu_j)| \, \|\mathbf{b}\|_2, \tag{5.20}$$

where $\kappa_2(W) = \|W\|_2\|W^{-1}\|_2$ is the condition number of the eigenvector matrix. As shown by Greenbaum et al. (1996) and discussed earlier $\kappa_2(W)$ can be so large that this bound is meaningless as a practical means of estimating $\|\mathbf{d}_k\|_2$.

The basic idea of the analysis used here can be found in Hackbusch (1994, Section 7.3.6). A similar analysis was used by van der Vorst and Vuik (1993) to study the superlinear convergence of GMRES.

Another typical observation is that GMRES converges in theory in not more than n iterations. This can be verified by setting $\Gamma = \mathbb{N}_n$ then $\|Q_\Gamma W^{-1}\mathbf{b}\|_2 = 0$ and hence $\|\mathbf{d}_k\|_2 = 0$ for $k \geq |\Gamma|$.

5.2.2 Some eigenvalue perturbation theory

In order to obtain a bound on the eigenvalues of $P_1^{-1}(A - \sigma^i M)P_2^{-1}$ we use a perturbation result which we then apply to our situation.

Theorem 5.7 *Let U be non-singular and ν be an eigenvalue of $D + E$ but not of D, then*

$$\|U(D - \nu I)^{-1}U^{-1}\|^{-1} \leq \|UEU^{-1}\| .$$

Proof: Proof see Stewart and Sun (1990, p.171). □

Corollary 5.8 *Let $D := P_1^{-1}(A - \lambda_1 M)P_2^{-1}$ be diagonalisable and denote its eigenvalue decomposition by $U\Lambda_D U^{-1}$. Further let $B^i := P_1^{-1}(A - \sigma^i M)P_2^{-1}$ have the eigenvalues μ_j^i with $|\mu_1^i| \leq |\mu_j^i|$ $\forall j$ while $C_5 := \|UP_1^{-1}MP_2^{-1}U^{-1}\|$. Then $|\mu_1^i| \leq |\lambda_1 - \sigma^i|\, C_5$.*
Proof: Set $E^i := B^i - D = (\lambda_1 - \sigma^i)P_1^{-1}MP_2^{-1}$, then $\|UE^iU^{-1}\| \leq C_5\, |\lambda_1 - \sigma^i|$. Now we use the fact that Λ_D is diagonal and μ_1^i the smallest eigenvalue of B^i, therefore

$$\|U(D - \mu_1^i I)^{-1}U^{-1}\|^{-1} = \|(\Lambda_D - \mu_1^i I)^{-1}\|^{-1} = |\mu_1^i|,$$

hence with Theorem 5.7 we obtain $|\mu_1^i| \leq C_5\, |\lambda_1 - \sigma^i|$. □

This proof makes use of the fact that μ_1^i is the smallest eigenvalue of the diagonalisable matrix D. If D is not diagonalisable, so Λ_D a Jordan matrix, then $\|\,(\Lambda_D - \mu_1^i I)^{-1}\,\|_2^{-1}$ is given by the smallest singular value of $\Lambda_D - \mu_1^i I$. As long as the smallest eigenvalue in modulo of $\Lambda_D - \mu^i I$ is non defective and well separated, the smallest singular value equals the absolute value of this eigenvalue. So, if $|\lambda - \sigma^i|$ is small enough and λ_1 a simple eigenvalue, then $|\mu_J^i|$ equals the smallest singular value. Earlier in Section 4.1 we derived a lower bound on the smallest eigenvalue of a Jordan matrix J, $\|J\| \geq \min_{\mu_j}\{|\mu_j| - \delta_j\}$ where $\{\mu_j\}_{j=1}^n$ are the eigenvalues of J and $\delta_j = 1$ if μ_j is defective and $\delta_j = 0$ otherwise. However, this bound is not of much use in case of preconditioned solves, as idealy the eigenvalues of the preconditioned system are

clustered around -1 and 1. Nevertheless, our practical experience is that the singular values corresponding to the Jordan blocks have no effect on the linear bound.

In the following we prove a lower bound on $|\mu_1^i|$. Therefore we use *gap* as defined in (4.19). We recall that $|\lambda_1 - \sigma^i| \leq \frac{1}{2} gap$ implies that $|\lambda_1 - \sigma^i|$ is the smallest singular value of $J - \sigma^i I$, see Section 4.1.

Lemma 5.9 *Let μ_1^i be the smallest eigenvalue of $P_1^{-1}(A - \sigma^i M)P_2^{-1}$ and $AV = MVJ$ the eigen-decomposition of the pencil (A, M). If $0 < |\lambda_1 - \sigma^i| \leq \frac{1}{2} gap$ then*

$$|\lambda_1 - \sigma^i| \|P_1\|^{-1} \|P_2\|^{-1} \|M^{-1}\|^{-1} \|V\|^{-1} \|V^{-1}\|^{-1} \leq |\mu_1^i|$$

Proof: From (4.7) we know that $(A - \sigma^i M)^{-1} M = V(\Lambda - \sigma^i I)V^{-1}$. Denote by ω_n the smallest singular value of $P_1^{-1}(A - \sigma^i M)P_2^{-1}$ then $\omega_n \leq |\mu|$. Using the fact that $\omega_n = \|P_1^{-1}(A - \sigma^i M)P_2^{-1}\|$ gives

$$\begin{aligned}
|\mu_1|^{-1} &\leq \omega_n^{-1} = \|P_2(A - \sigma^i M)^{-1} P_1\| \\
&\leq \|P_2\| \|P_1\| \|(A - \sigma^i M)^{-1}\| \\
&\leq \|P_2\| \|P_1\| \|(A - \sigma^i M)^{-1} M M^{-1}\| \\
&\leq \|P_2\| \|P_1\| \|M^{-1}\| \|(A - \sigma^i M)^{-1} M\| \\
&\leq \|P_2\| \|P_1\| \|M^{-1}\| \|V(J - \sigma^i I)^{-1} V^{-1}\| \\
&\leq \|P_2\| \|P_1\| \|M^{-1}\| \|V\| \|V^{-1}\| \|(J - \sigma^i I)^{-1}\| .
\end{aligned}$$

We conclude the proof by using that $|\lambda_1 - \sigma^i| \leq \frac{1}{2}$ implies that $|\lambda_1 - \sigma^i|$ is the smallest singular value of $J - \sigma^i I$. □

Stewart and Sun (1990, Chapter 4, Theorem 1.1) provides the fact that eigenvalues are continuous functions of the matrix entries, i.e. $\mu_j^i = \mu_j(\sigma^i)$ is continuous. Now given a set $\Gamma \subset \mathbb{N}_n$ then denote

$$D_\Gamma := \{\mu_j | j \notin \Gamma,\ \mu_j = \mu_j(\sigma) \text{ with } |\lambda_1 - \sigma| \leq \frac{1}{2} gap\}. \tag{5.21}$$

We observe that D_Γ is compact as $\{\sigma |\ |\lambda_1 - \sigma| \leq \frac{1}{2} gap\}$ is compact and $\mu_j(\sigma)$ continuous. Next we prove that $D_\Gamma \subset \mathbb{C}$ is a compact set and $0 \notin D_\Gamma$.

Lemma 5.10 *Let the eigenvalues of B^i be ordered such that $|\mu_1^i| \leq |\mu_2^i| \leq \ldots \leq |\mu_n^i|$. If $|\lambda_1 - \sigma^i| \leq C_9 gap$ where $0 < C_9 < 1$ then $\exists C_7, C_8 > 0$ such that $C_7 \leq |\mu_2^i| \leq |\mu_n^i| \leq C_8$.*

Proof: The arguments we applied to D_Γ are also valid for

$$\Omega := \bigcup_{j=2}^{n} \mu_j(\{\sigma |, |\lambda_1 - \sigma| \leq C_9 gap\})$$

and hence Ω is compact. Therefore $\exists C_8 > 0$ such that $|\mu_j^i| \leq C_8$. As $\mu(\sigma)$ eigenvalue of

$P_1^{-1}(A-\sigma M)P_2^{-1}$ and P_1 and P_2 non-singular, $\mu_j(\sigma)=0$ only if $A-\sigma M$ is singular, which is the same as σ being an eigenvalue of $A\mathbf{x}=\lambda M\mathbf{x}$. Now $|\lambda_1-\sigma|\leq \frac{1}{2}gap$ contains only one σ which is an eigenvalue of $A\mathbf{x}=\lambda M\mathbf{x}$, that is $\sigma=\lambda_1$. For $\sigma=\lambda_1$ we have $\mu_1=0$ and $\mu_j\neq 0$ for $j\geq 2$, therefore $0\notin\Omega$. □

As $D_\Gamma\subset\Omega$ we have D_Γ compact and $0\notin D_\Gamma$, so η_Γ^k is well defined.

5.2.3 Preconditioned GMRES as linear solver for shifted linear systems

In Chapter 6 we apply GMRES to sequences of shifted linear systems. Further we consider the cases where either unpreconditioned GMRES or preconditioned GMRES is applied to such a sequence. To simplify later analysis we discuss both cases here and present a Corollary to Lemma 5.6 applicable to both cases. In case of preconditioned GMRES we might apply a left or a right preconditioner or both to $(A-\sigma^i M)\mathbf{y}^i=\mathbf{b}^i$. However in all cases the preconditioned system can be written as

$$P_1^{-1}(A-\sigma^i M)P_2^{-1}\mathbf{z}^i \;=\; P_1^{-1}\mathbf{b}^i, \tag{5.22}$$

with $\mathbf{y}^i=P_2^{-1}\mathbf{z}^i$. We define the residual for the ith linear system by

$$\mathbf{d}_k^i \;:=\; P_1^{-1}\mathbf{b}^i - P_1^{-1}(A-\sigma^i M)P_2^{-1}\mathbf{z}_k^i, \tag{5.23}$$

where $\mathbf{z}_k^i$ is the solution after k GMRES iterations on the ith linear system.

Corollary 5.11 *Consider GMRES being applied to the linear systems* $P_1^{-1}(A-\sigma^i M)P_2^{-1}\mathbf{z}^i = P_1^{-1}\mathbf{b}^i$. *Denote the eigenvalues of* $P_1^{-1}(A-\sigma^i M)P_2^{-1}$ *by* $\mu_1^i,\ldots,\mu_n^i$ *and those of* $M^{-1}A$ *by* $\lambda_1,\ldots,\lambda_n$. *Assume* $0<|\lambda_1-\sigma^i|<\frac{1}{2}gap$, P_1 *and* P_2 *non-singular and* $\Gamma\subset\mathbb{N}_n$ *such that* $\{j|\mu_j defective\}\subset\Gamma$ *and* $\eta_{D_\Gamma}^k\to 0$, *where* D_Γ *is defined in (5.21). Then GMRES converges and there exists* $p>0$ *and* $q\in(0,1)$ *as given by Lemma 5.5 such that for the residual as defined in (5.23) the bound*

$$\|\mathbf{d}_k^i\|_2 \;\leq\; (q_\Gamma)^{k-|\Gamma|}p_\Gamma\,|\lambda_1-\sigma^i|^{-1}\,\chi^i \tag{5.24}$$

holds for all i *and* $k\geq|\Gamma|$. *Here* $q_\Gamma=q$ *and*

$$p_\Gamma \;=\; pC_6\max_{\xi\in D_\Gamma}\left((C_8+|\xi|)\prod_{j\in\Gamma\setminus\{1\}}\frac{C_8+|\xi|}{C_2}\right), \tag{5.25}$$

where $|\lambda_1-\sigma|\,/C_6\leq|\mu_1|$ *as in Corollary 5.9 while* $\chi^i:=\|W^iQ_\Gamma\|_2\|Q_\Gamma(W^i)^{-1}P_1^{-1}\mathbf{b}^i\|_2$, *with* W^i *being the matrix of eigenvectors of* $P_1^{-1}(A-\sigma^i M)P_2^{-1}$.

Proof: For each system the conditions of Lemma 5.6 are satisfied and to bound $|\mu_1^i|^{-1}$

we use Lemma 5.9. Further, to bound $\widetilde{p_\Gamma{}^i}$ as in Lemma 5.6 we use Lemma 5.10 and thus (5.25) holds. □

The constants p_Γ and q_Γ might be improved using $D_\Gamma^i := \{\mu_j^i | j \notin \Gamma\}$, however later it is more convenient to have p_Γ and q_Γ independent of σ^i.

A similar result to Corollary 5.11 has been obtained by Campbell et al. (1996), however their result lacks the projection term χ^i, which will be a key quantity in Chapter 6.

5.3 Literature

Here we present a brief overview of literature on the convergence of GMRES. Further we discuss a few articles concerned with enhancements of GMRES, namely, augmenting and restarting.

A polynomial convergence bound for GMRES like Lemma 5.6 but with $\Gamma = \{\}$ has been published by Saad and Schultz (1986) together with the algorithm. Other standard sources for the algorithm and basic results are for example Kelley (1995); Saad (1996) and Greenbaum (1997).

Greenbaum et al. (1996) showed that any non-increasing convergence curve is possible for the GMRES residual independent of the spectrum of the system matrix. As we explained in the introduction this effect is often achieved by deterioration of the conditioning of the eigenvectors.

While the system matrix B is normally assumed to be nonsingular, Brown and Walker (1997) show that GMRES finds the least squares solution or an approximation to it in the singular case. The authors also discuss the case of nearly singular systems. One of their findings is that GMRES applied to nearly singular systems might behave almost as in the singular case. Based on their analysis and their computational results it would appear that GMRES is not a good solver for inexact inverse iteration. However their observations are for the general case and do not take into account that the right-hand side is somehow special in inexact inverse iteration.

The analysis from van der Vorst and Vuik (1993) gives considerably more insight for our application. In practice GMRES often shows superlinear convergence behaviour. The aim of the analysis of van der Vorst and Vuik (1993) is to provide an understanding of such behaviour. As they show, a key factor is how well small eigenvalues of the system matrix B are approximated by the Ritz values of A with respect to the subspace used by GMRES. Once such eigenvalues are approximated well enough the rate of convergence improves as if these values were not present. For our application there are two situations of further interest. For the first we assume that the right hand side is an approximation of the eigenvector corresponding to the eigenvalue smallest in magnitude of the (preconditioned) system. In this case GMRES detects this almost

zero eigenvalue quickly, perhaps in one iteration, depending on the quality of the right-hand side. Then, according to van der Vorst and Vuik (1993), the further convergence is almost not affected by this eigenvalue. In the other case, when the right-hand side is not an approximation of the eigenvector then the GMRES algorithm might need considerable effort to find an approximation to this eigenvector. This observation is also supported by the analysis of Ipsen (1998a) for the case $B = B^T$.

A comparison of GMRES with other iterative techniques for non-symmetric systems had been published by Nachtigal et al. (1992). Basically all iterative methods for non-symmetric matrices can fail in practice and can be outclassed by other methods so that there is no 'best' solver.

In practice a method called restarted GMRES is used frequently. As GMRES needs to store the subspace vectors the storage requirement increases with the number of iterations. Due to limited storage one might restrict the Krylov subspace size. This implies that the number of GMRES iterations will be limited and it might happen that GMRES does not converge sufficiently inside this limited number of iterations. Restarted GMRES uses restarts of GMRES with the previous residual as the new right-hand side. A major drawback of restarted GMRES is the possible lack of convergence; there are examples where restarted GMRES with restarts every twenty iterations stagnates, so does not convergence, while GMRES with restarts every second iteration converges, for more see Embree (2003). In this thesis we do not consider restarted GMRES, due to such convergence problems. Using the approximate solution vectors to construct another Krylov subspace leads to flexible (restarted) GMRES, see Saad (1993). However, to overcome stagnation Simoncini and Szyld (2002) suggest an inner-outer GMRES algorithm, where size of the inner Krylov spaces must be non decreasing.

Another approach to overcome poor convergence is to use augmented GMRES. The idea of augmenting GMRES is to provide information of the eigenvalues close to zero, which might enhance the convergence speed as explained earlier. In practice this approach might be combined with restarted GMRES, where it might improve the convergence, but does not resolve the problem of stagnation. Here we do not consider augmented GMRES, but point out that our analysis might be extended to cover this case. For more on augmenting see Chapman and Saad (1997) and for combining restarting and deflating see Morgan (1995).

Chapter 6

Efficient Variations of Inexact Inverse Iteration using GMRES for the GEP

As in Chapter 4 we consider the generalised unsymmetric eigenvalue problem, (GEP),

$$A\mathbf{x} = \lambda M\mathbf{x}, \tag{6.1}$$

with $A, M \in \mathbb{R}^{n\times n}$ and M spd, where the eigenpair $(\lambda, \mathbf{x}) = (\lambda_1, \mathbf{v}_1)$ is sought. In Chapter 4 we considered the convergence of inexact inverse iteration independent of any particular solver. In order to analyse the efficiency of inexact inverse iteration as an inner-outer type algorithm we have to consider a specific solver or a class of solvers. Here we assume that GMRES is applied to the arising linear systems.

Earlier in Chapter 3 we considered the efficiency of inexact inverse iteration applied to the standard symmetric eigenvalue problem and using MINRES. In this chapter we now adapt the key ideas and the analysis of Chapter 3 to the GEP. However the form of presentation will be changed.

Some of the methods discussed in Chapter 4 use a shift converging towards the desired eigenvalue. When the shift converges to the sought eigenvalue the linear systems get harder to solve and the systems get closer to being singular. The application of GMRES to singular and nearly singular systems was studied in Brown and Walker (1997). An almost singular system can cause problems especially if a good solution of the system is needed, that is the error in the solution should be small. According to our experience these problems are largely due to round-off errors which are less fatal for MINRES. However we are not primarily interested in solving the linear system but gaining a good approximation of the sought eigenvector. As the analysis of van der Vorst and Vuik (1993) shows the convergence of GMRES is not hampered too much by a few critical eigenvalues. We explored this in our convergence analysis for GMRES in

Chapter 5. From our efficiency analysis we will observe that when shifting towards the sought eigenvalue the cost increase for a linear solve does not outweight the benefits in terms of a better approximation to the sought eigenvalue.

We start, in Section 6.2, by defining the number of GMRES iterations, as a measure for the cost of a linear solve. Based on this we define the cost of calculating an approximate eigenpair $(\varrho^i, \mathbf{x}^i)$ of the GEP (6.1) by the sum of all GMRES iterations.

Now, as in Chapter 3, we use in Section 6.2 a convergence bound for the solver to derive an a-posteriori upper bound on the number of inner-iterations. This upper bound is a key result to understand the efficiency of the methods discussed earlier in Section 4.3. In Section 6.3 we discuss this first result with respect to those methods. Our second key result with respect to the efficiency of inexact inverse iteration will be presented in Section 6.4. This result provides a-posteriori upper bounds on the total number of inner-iterations. Based on this result we observe that from the methods considered in Section 4.3, RQId, that is an inexact RQ iteration with decreasing tolerance, and InvitWd, that is inexact inverse iteration using the Wilkinson update and a decreasing tolerance, are the most efficient.

Finally in Section 6.5 we provide numerical results to support our Theory.

6.1 Costs

As in Chapter 3 we refer to the GMRES iterations as inner-iterations in contrast to the iterations of inexact inverse iteration, referred to as outer-iterations. In contrast to Chapter 3 and MINRES the cost of a linear solve using GMRES is not necessarily linear in the number of GMRES iterations, as we explain in the following.

The major cost terms in GMRES are preconditioned matrix vector products, storage and orthonormalisation. While the cost of orthonormalising the Krylov basis vectors grows quadratically in the number of inner-iterations k, the other two major costs are linear in k. The linearity of the storage requirement with respect to k is a problem as the storage available is usually limited, which in turn restricts the number of inner-iterations which can be performed. However, if enough storage is available and the cost for orthonormalisation is small then the main cost of a solve using GMRES is given by the number of matrix vector products. In practice this assumption is for large sparse systems not unreasonable if very good and often expensive preconditioners are applied.

We are aware that the cost of a preconditioned matrix vector product is not fixed for all preconditioners, however, here we only consider the case where the cost of a preconditioned matrix vector product is independent of the vector applied to. Hence the only relevant cost remaining is the number of preconditioned matrix vector products, which equals the number of inner iterations.

Definition 6.1 *Apply GMRES to* $B\mathbf{y} = \mathbf{b}$ *using the accuracy requirement* $\|\mathbf{res}\| \leq \tau \|\mathbf{b}\|$ *as stopping condition. Then define* $\mathcal{L} \in \mathbb{N}$ *as the minimal number of iterations required by GMRES such that the residual condition is achieved, that is*

$$\|\mathbf{res}_{\mathcal{L}}\|_2 \leq \tau \|\mathbf{b}\|_2 \quad \text{and} \quad \|\mathbf{res}_k\|_2 > \tau \|\mathbf{b}\|_2 \quad \forall 0 \leq k < \mathcal{L}.$$

Again our prime interest is to control the sum of matrix-vector multiplications, which equals the total number of inner-iterations. Hence we use the total number of inner-iterations as a measure for the overall cost. Therefore we rewrite Definition 3.9 for the GEP.

Definition 6.2 *Given matrices* $A, M \in \mathbb{R}^{n\times n}$, *where* M *is spd, a starting vector* $\mathbf{x}^0$, *a sequence of shifts* σ^i *and a sequence of accuracy requirements* τ^i *for the linear solves. Then define the total cost* $\mathcal{T}$ *as the sum of all inner-iterations used to achieve a generalised tangent* $t \leq t^*$, *that is* $\mathcal{T} := \sum_{i=0}^{\mathcal{N}} \mathcal{L}^i$, *where* $\mathcal{N}$ *denotes the number of outer-iterations performed.*

6.2 Efficiency analysis

We start this section with a short summary of the notation used in Chapters 4 and 5 with respect to our later needs. In Section 6.2.2, we present Theorem 6.3, our first key result concerning the number of inner iteration per outer iteration, $\mathcal{L}^i$.

6.2.1 Notation

As in Chapter 4 we consider the generalised eigenvalue problem $A\mathbf{x} = \lambda M\mathbf{x}$, where $A, M \in \mathbb{C}^{n\times n}$ and M spd. We assume that λ_1, the sought eigenvalue is simple and that *gap*, as defined in (4.19), is positive. The sought eigenvector is denoted by $\mathbf{v}_1^R$ and for its approximation, $\mathbf{x}^i$, we consider the splitting $\mathbf{x}^i = \alpha^i(c^i\mathbf{v}_1^R + s^i\mathbf{u}^i)$, see (4.9), which gives the generalised tangent $t^i = |s^i| \,/\, |c^i|$, see (4.10). Further let V_L denote the matrix of left eigenvectors and let $\mathbf{v}_1^L = V_L\mathbf{e}_1$ be the left eigenvector corresponding to λ_1. We denote the residuals for the linear systems arising in inexact inverse iteration, see Algorithm 5 (page 89), by

$$\mathbf{res}_k^i = \mathbf{b}^i - (A - \sigma^i M)\mathbf{y}_k^i.$$

However depending on the actual method, the linear system being solved might differ, nevertheless we can write it as

$$P_1^{-1}(A - \sigma^i M)P_2^{-1}\mathbf{z}_k^i \;=\; P_1^{-1}B^i\mathbf{b}^i,$$

	σ^i	B^i	$\mathbf{b}^i$
InvitFd	σ^0	I	$M\mathbf{x}^i$
RQIf	ϱ^i	I	$M\mathbf{x}^i$
RQId	ϱ^i	I	$M\mathbf{x}^i$
InvitWf	ϱ^i_W	I	$M\mathbf{x}^i$
InvitWd	ϱ^i_W	I	$M\mathbf{x}^i$
PInvit	ϱ^i	I	$P\mathbf{x}^i$
GY	σ^0	$I-\varphi^i(AM^{-1}-\sigma^0 I)$	$M\mathbf{x}^i$
GRQIf	ϱ^i_G		$M\mathbf{x}^i$
PInvitGRQ	ϱ^i_G		$P\mathbf{x}^i$

Table 6.1: Practical methods as discussed in Chapter 4

where $P = P_1P_2$ is the preconditioner used implicitly in GMRES. As GMRES is an iterative solver we write for the kth GMRES iterate $\mathbf{z}^i_k$, approximating $\mathbf{z}^i$. The resulting residual we denote as in (5.23) by

$$\mathbf{d}^i_k \;:=\; P_1^{-1}B^i\mathbf{b}^i - P_1^{-1}(A-\sigma^i M)P_2^{-1}\mathbf{z}^i_k.$$

To relate the two linear systems for each method, Table 6.1 shows the specific choices for σ^i, B^i and $\mathbf{b}^i$ made for the methods discussed in Chapter 4. As indicated in Table 6.1 solving the standard system $(A-\sigma^i M)\mathbf{y}^i = M\mathbf{x}^i$, as for example, in InvitFd, leads to $B^i = I$ and $\mathbf{b}^i = M\mathbf{x}^i$. Also for PInvit, using the modified right-hand side, the choices $B^i = I$ and $\mathbf{b}^i = P\mathbf{x}^i$ for PInvit are obvious. In Chapter 4, page 104, we showed that GY can be viewed as inexact inverse iteration using a fixed shift and decreasing tolerance, hence we use $\mathbf{b}^i = M\mathbf{x}^i$. As the right-hand side for the actual solve is

$$\mathbf{r}^i_{GY} = M\mathbf{x}^i - (A-\sigma^0 M)\mathbf{y}^i \;=\; \Big(I-\varphi^i(AM^{-1}-\sigma^i I)\Big)M\mathbf{x}^i,$$

see (4.41), we set $B^i = I-\varphi^i(AM^{-1}-\sigma^i I)$. For all these methods the two residuals $\mathbf{res}^i_k$ and $\mathbf{d}^i_k$ satisfy

$$\mathbf{res}^i_k \;=\; P_1\mathbf{d}^i_k \tag{6.2}$$

Since we shall use the convergence bound for GMRES as stated in Corollary 5.11 we recall the definition of the terms used there. For a given index set $\Gamma \subset \mathbb{N}_n$ we define the projection matrix Q_Γ as in (5.18) by $Q_\Gamma = \mathrm{diag}(\delta_1(\Gamma),\ldots,\delta_n(\Gamma))$ with $\delta_j(\Gamma) = 1$ if $j \notin \Gamma$ and $\delta_j(\Gamma) = 0$ otherwise.

Let $D \subset \mathbb{C}$ then we defined in (5.4)

$$\eta^k_D \;=\; \min_{f\in\Pi^1_k}\max_{z\in D}|f(z)|,$$

where Π_k^1 denotes the set of polynomials with degree $\leq k$ and $f(0) = 1$.

Finally we denote the eigenvalues of $P_1^{-1}(A - \sigma^i M)P_2^{-1}$ by $\mu_1^i, \ldots, \mu_n^i$ where μ_1^i is the eigenvalue with smallest modulus.

6.2.2 Cost per outer iteration

We now present our first key result concerning the efficiency of inexact inverse iteration using unpreconditioned or preconditioned GMRES for the arising linear systems.

Theorem 6.3 *Let Inexact Inverse Iteration, Algorithm 5, with B^i and (preconditioned) GMRES as linear solver be applied to $A, M \in \mathbb{R}^{n\times n}$, where M is spd. Further let λ_1 be a simple eigenvalue of the matrix pair A, M. Assume that the preconditioner $P = P_1P_2$ is non-singular and that GMRES is applied to the linear systems are of the form $(A-\sigma^i M)\mathbf{z}^i = B^i\mathbf{b}^i$. Also let $\Gamma \subset \mathbb{N}_n$ with $1 \in \Gamma$ and all $j \in \Gamma$ where μ_j^i defective for some i, and assume $D \supset \{\mu_j^i | j \notin \Gamma, i = 0, 1, 2, \ldots\}$ such that $\eta_D^k \to 0$, where η_D^k is defined in (5.4). If the conditions of Theorem 4.2 are satisfied then $t^i \to 0$. Further, if additionally there exists $C_4 > 0$ such that*

$$\|(I - \mathbf{e}_1\mathbf{e}_1^T)V_L^H\mathbf{b}^i\| \;\leq\; C_4\,\|V_L^H\mathbf{res}^i\|, \tag{6.3}$$

then there exist $p_\Gamma > 0$, $q_\Gamma \in (0,1)$ such that

$$\mathcal{L}^i \;\leq\; 1 + |\Gamma| + \log\left(C_5\frac{\|W^iQ_\Gamma\|\,\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i\|}{|(\mathbf{v}_1^L)^H\mathbf{b}^i|\;t^{i+1}}\right)/\log((q_\Gamma)^{-1}), \tag{6.4}$$

where $C_5 := 2(1+C_4)p_\Gamma\,\|P_1\|\|V_L\|\,(gap(1-C_3))^{-1}$. The matrix W^i denotes the matrix of eigenvectors of the preconditioned system matrix $P_1^{-1}(A-\sigma^iM)P_2^{-1}$.

Proof: The conditions of Theorem 4.2 imply $0 < |\lambda_1 - \sigma^i| \leq \frac{1}{2}gap$, whereby the conditions of Corollary 5.11 are satisfied. Hence GMRES converges, i.e. $\mathbf{d}_k^i \to 0$ for $k \to \infty$ and thereby $\mathbf{res}_k^i = P_1\mathbf{d}_k^i \to 0$. As $\mathbf{res}_k^i \to 0$ there exists $\mathcal{L}^i$ as in Definition 6.1 and $\|\mathbf{res}_{\mathcal{L}^i}^i\| \leq \tau^i$. Now applying Theorem 4.2 we obtain convergence for inexact inverse iteration.

As $\mathcal{L}^i$ exists Definition 6.1 gives

$$\|\mathbf{res}_{\mathcal{L}^i}^i\| \;\leq\; \tau^i \;<\; \|\mathbf{res}_{\mathcal{L}^i-1}^i\|\,.$$

Combining this with Corollary 5.11 we obtain

$$\begin{aligned}\tau^i \;&<\; \|\mathbf{res}_{\mathcal{L}^i-1}^i\| \;\leq\; \|P_1\|\|\mathbf{d}_{\mathcal{L}^i-1}^i\| \\ &\leq\; \|P_1\|\,(q_\Gamma)^{\mathcal{L}^i-1-|\Gamma|}p_\Gamma\,|\lambda_1 - \sigma^i|^{-1}\,\chi^i,\end{aligned}$$

where $\chi^i = \|W^i Q_\Gamma\| \|Q_\Gamma (W^i)^{-1} P_1^{-1} B^i \mathbf{b}^i\|$. We solve for $\mathcal{L}^i$ to obtain

$$\mathcal{L}^i \leq 1 + |\Gamma| + \log\left(\frac{\|P_1\| \; p_\Gamma \, \chi^i}{|\lambda_1 - \sigma^i| \; \tau^i}\right) / \log((q_\Gamma)^{-1}). \tag{6.5}$$

To link this with the outer convergence we use the one-step bound for the generalised tangent, (4.28), which we rearrange as

$$\begin{aligned}
& t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{gap - |\lambda_1 - \sigma^i|} \frac{\|(I - \mathbf{e}_1 \mathbf{e}_1^T) V_L^H \mathbf{b}^i\| + \|V_L^H \mathbf{res}^i\|}{|(\mathbf{v}_1^L)^H \mathbf{b}^i| - \|V_L^H \mathbf{res}^i\|} \\
\Longrightarrow \quad & t^{i+1} \leq 2 \frac{|\lambda_1 - \sigma^i|}{gap} \frac{(1 + C_4) \, \|V_L^H \mathbf{res}^i\|}{(1 - C_3) \, |(\mathbf{v}_1^L)^H \mathbf{b}^i|} \\
\Longrightarrow \quad & t^{i+1} \leq 2 \frac{|\lambda_1 - \sigma^i|}{gap} \frac{(1 + C_4) \, \|V_L^H\|}{(1 - C_3) \, |(\mathbf{v}_1^L)^H \mathbf{b}^i|} \, \tau^i \\
\Longleftrightarrow \quad & \frac{1}{|\lambda_1 - \sigma^i| \, \tau^i} \leq \frac{2 \, (1 + C_4) \, \|V_L^H\|}{gap \, (1 - C_3) \, |(\mathbf{v}_1^L)^H \mathbf{b}^i|} \frac{1}{t^{i+1}}.
\end{aligned} \tag{6.6}$$

We conclude the proof by inserting the last inequality into the bound on $\mathcal{L}^i$, (6.5), to obtain

$$\begin{aligned}
\mathcal{L}^i \leq \; & 1 + |\Gamma| + \log\left(\frac{2(1 + C_4) p_\Gamma \, \|P_1\| \, \|V_L\|}{gap(1 - C_3)}\right) / \log((q_\Gamma)^{-1}) \\
& + \log\left(\frac{\chi^i}{t^{i+1} \, |(\mathbf{v}_1^L)^H \mathbf{b}^i|}\right) / \log((q_\Gamma)^{-1})
\end{aligned}$$

and by using the definition of C_5 we gain (6.4). □

Obviously with $\Gamma = \mathbb{N}_n$, then $\|W^i Q_\Gamma\| = 0$, and hence $\mathcal{L}^i < 1 + n$. However we are interested in the case where Γ is a small index set, $|\Gamma| \ll n$, but then the contribution of the log-terms is not negligible.

Comparing Theorem 6.3 with Theorem 4.2, we observe to conditions on $\|V_L^H \mathbf{res}^i\|$. In Theorem 6.3 part c) we have $\|V_L^H \mathbf{res}^i\| \leq C_3 \, |s^i|^{\gamma_2} |(\mathbf{v}_1^L)^H \mathbf{b}^i|$, so lowering C_3 implies a smaller residual. Further, Remark 4.4 shows that lowering C_3 improves the rate of convergence. Now in Theorem 4.2 we asked for an upper bound on $\| V_L^H \mathbf{res}^i \|$. A smaller residual $\|V_L^H \mathbf{res}^i\|$ implies a relaxation of the lower bound, which in turn forces an increase in C_4. As $\mathcal{L}^i \propto \log(1 + C_4)$, reducing C_3 leads to an increase in $\mathcal{L}^i$.

Remark 6.4 *We expect the number of inner-iterations to increase when the tolerance condition in Theorem 4.2 is tightened.*

We observe this effect in practice for all methods, but we provide only for InvitFd numerical results, see Table 6.2. While reducing C_3 leads to better outer convergence, it is so far not clear if increasing or decreasing C_3 improves the overall performance.

Further, we point out that the bound (6.4) is independent of the size of the matrices A, M, and depends only on the distribution of the eigenvalues.

6.3 Practical Methods

Before we discuss the implications of Theorem 6.3 for each of the practical methods, we prepare a few tools for that task. These tools are mainly about deriving a bound for $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i\|$.

6.3.1 Some bounds

For GY we have $B^i\mathbf{b}^i \to 0$ and hence $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i\| \to 0$.

However, for the other methods the situation is more subtle.

Under the conditions of Theorem 6.3, t^i tends to zero and also the sequences $(B^i\mathbf{b}^i)$, $(\mathbf{b}^i)$, (α^i) and (W^i) converge to a limit. For our later convenience we denote these limits with subindex 0, so $\lim_{i\to\infty} W^i = W_0$, similar for the other sequences.

As $B^i\mathbf{b}^i$ is effected by the scaling α^i which is implicit in $\mathbf{x}^i$ we consider from now on $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\|$. Further we use the fact that for all $\eta \in \mathbb{C}$

$$\begin{aligned} & \|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \\ \leq\ & \|Q_\Gamma(W^i)^{-1}W^i\mathbf{e}_1\eta\| + |\eta|\,\|Q_\Gamma(W^i)^{-1}\|\,\|W^i\mathbf{e}_1 - W_0\mathbf{e}_1\| \\ & + \|Q_\Gamma(W^i)^{-1}\|\,\|\eta W_0\mathbf{e}_1 - P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1}\| \\ & + \|Q_\Gamma(W^i)^{-1}\|\,\|P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1} - P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\|\,. \end{aligned} \tag{6.7}$$

As $1 \in \Gamma$ we have $\|Q_\Gamma(W^i)^{-1}W^i\mathbf{e}_1\| = 0$. Now we define $C_6 := \max_i \|Q_\Gamma(W^i)^{-1}\|$, then for $|\Gamma| < n$ we observe that $C_6 > 0$. So, we derive from (6.7) the inequality

$$\begin{aligned} \|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\|/C_6 \ \leq\ & |\eta|\,\|W^i\mathbf{e}_1 - W_0\mathbf{e}_1\| \\ & + \|\eta W_0\mathbf{e}_1 - P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1}\| \\ & + \|P_1^{-1}(B_0\mathbf{b}_0(\alpha_0)^{-1} - B^i\mathbf{b}^i(\alpha^i)^{-1})\|, \end{aligned} \tag{6.8}$$

which holds for all $\eta \in \mathbb{C}$. In order to bound the first term on the right-hand side in (6.8) we use an eigenvector perturbation type result from Chatelin (1993).

Lemma 6.5 *Given the matrices G and $E \in \mathbb{C}^{n\times n}$. Let $\zeta \in \mathbb{C}$ and let $\mathbf{z}_\zeta$ denote the eigenvector of $G + \zeta E$ corresponding the eigenvalue μ_ζ. Then*

$$\|\mathbf{z}_0 - \mathbf{z}_\zeta\| \ =\ \|\Sigma^\perp E\mathbf{z}_0\|\,|\zeta| + O(|\zeta|^2),$$

where $\Sigma^\perp := U(U^HGU - \mu_0 I)U^H$ with $(\mathbf{z}_0, U)$ unitary basis of $\mathbb{C}^n$.

Proof: See Proposition 4.3.1 and corresponding proof in Chatelin (1993). □

Corollary 6.6 *Assume the conditions of Theorem 6.3 are satisfied. If* $|\lambda_1 - \sigma^i| \leq C_1 t^i$ *then there exists* $C_7 > 0$ *with*

$$\|W^i \mathbf{e}_1 - W_0 \mathbf{e}_1\| \;\leq\; C_7\, t^i. \tag{6.9}$$

Proof: Under the conditions of Theorem 6.3, t^i tends to zero and thus $\sigma^i \to \lambda_1$. As a result the matrix W_0 is the eigenvector matrix of $G := P_1^{-1}(A - \lambda_1 M)P_2^{-1}$. Now let $E := P_1^{-1} M P_2^{-1}$ and $\zeta^i := \lambda_1 - \sigma^i$, then W^i is the eigenvector matrix of $G - \zeta^i E$. Applying Lemma 6.5 we have

$$\|W^i \mathbf{e}_1 - W_0 \mathbf{e}_1\| \;\leq\; \|\Sigma^{\perp} E W_0 \mathbf{e}_1\| |\lambda_1 - \sigma^i| + O(|\lambda_1 - \sigma^i|^2).$$

Under the conditions of Theorem 6.3 we have $t^i \leq \frac{1}{4}$, hence there exists $C_7 > 0$ such that (6.9) holds. □

Next we provide a bound for the third term on the right-hand side of (6.8).

Lemma 6.7 *Assume the conditions of Theorem 6.3 are satisfied, then there exists* $C_8 > 0$ *such that*

$$\|P_1^{-1} B_0 \mathbf{b}_0 (\alpha_0)^{-1} - P_1^{-1} B^i \mathbf{b}^i (\alpha^i)^{-1}\| \;\leq\; C_8\, t^i.$$

Proof: As for GY $B_0 \mathbf{b}_0 = \mathbf{0}$ and $\|B^i \mathbf{b}^i\| = \|\mathbf{r}^i\| \leq t^i \mathcal{R}^*$ the result is obviously valid for these two methods. For the methods using the standard right-hand side, namely, InvitFd, RQIf, RQId, InvitWf and InvitWd, we have $B_0 \mathbf{b}_0 (\alpha_0)^{-1} = M \mathbf{v}_1^R$. Now

$$\begin{aligned} \|P_1^{-1}(M\mathbf{v}_1^R - M\mathbf{x}^i(\alpha^i)^{-1})\| \;&\leq\; \|P_1^{-1} M((1-c^i)\mathbf{v}_1^R + s^i \mathbf{u}^i)\| \\ &\leq\; t^i \Big(\|P_1^{-1} M \mathbf{v}_1^R\| + \|P_1^{-1} M \mathbf{u}^i\| \Big) \;\leq\; C_8\, t^i. \end{aligned}$$

Finally for PInvit, using the modified right-hand side, we have $P_1^{-1} B_0 \mathbf{b}_0 = P_2 \mathbf{v}_1^R$ and hence

$$\|P_2 \mathbf{v}_1^R - P_2 \mathbf{x}^i (\alpha^i)^{-1})\| \;\leq\; \|P_2((1-c^i)\mathbf{v}_1^R + s^i \mathbf{u}^i)\| \;\leq\; C_8\, t^i.$$

□

Another useful observation is that for all methods in Table 6.1 $|(\mathbf{v}_1^L)^H \mathbf{b}^i| \;/\; |\alpha^i|$ tends to some positive constant. Therefore, there exists $C_{10} > 0$ such that

$$\|W^i Q_r\|\, |\alpha^i| \;/\; |(\mathbf{v}_1^L)^H \mathbf{b}^i| \;\leq\; C_{10}. \tag{6.10}$$

6.3.2 Practical Methods and cost per iteration

Now we are fully equipped to bound $\|Q_r(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i\|$ and the following Lemma provides the sought bounds. This Lemma will enable us to discuss the implications of Theorem 6.3 for each method with respect to the number of inner-iterations per outer iteration, $\mathcal{L}^i$.

Lemma 6.8 *Assume the conditions of Theorem 6.3 being satisfied. Denote the matrix of eigenvectors of $P_1^{-1}(A-\sigma^i M)P_2^{-1}$ by W^i and let $W^i\mathbf{e}_1$ be the eigenvector corresponding to the eigenvalue with smallest modulus. Further let $\mathbf{x}^i = \alpha^i(c^i\mathbf{v}_1^R + s^i\mathbf{u}^i)$ be an approximation of $\mathbf{v}_1^R$ and consider B^i and $\mathbf{b}^i$ as given in Table 6.1.*

a) If $\angle(M\mathbf{v}_1^R, \mathbf{v}_1^R) = 0$ then for the methods InvitFd, RQIf, RQId, InvitWf and InvitWd using unpreconditioned GMRES there exists a constant $C_9 > 0$ with

$$\|Q_r(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \;\leq\; C_9\, t^i. \tag{6.11}$$

b) If $\angle(P_2\mathbf{v}_1^R, P_1^{-1}M\mathbf{v}_1^R) = 0$ then for the methods InvitFd, RQIf, RQId, InvitWf and InvitWd using preconditioned GMRES there exists $C_9 > 0$ such that (6.11) holds.

c) For the method PInvit there exists a constant $C_9 > 0$ such that (6.11) holds.

d) For the method GY there exists a constant $C_9 > 0$ such that (6.11) holds.

Proof: To prove parts a) and b) for InvitFd we observe that $W^i = W_0$, and there exists η with $\eta P_2\mathbf{v}_1^R = P_1^{-1}M\mathbf{v}_1^R$ hence $W_0\mathbf{e}_1 = P_2\mathbf{v}_1^R$ as

$$P_1^{-1}(A-\sigma^0 M)P_2^{-1}P_2\mathbf{v}_1^R \;=\; (\lambda_1-\sigma^i)P_1^{-1}M\mathbf{v}_1^R \;=\; \eta P_2\mathbf{v}_1^R.$$

Using that the first two terms in inequality (6.8) are zero we conclude the proof of parts a) and b) for InvitFd by using Lemma 6.7 and set $C_9 = C_{10}$.

For the remaining methods in parts a) and b) and for part c) we use inequality (6.8) together with Corollary 6.6 and Lemma 6.7 to obtain

$$\begin{aligned}&\|Q_r(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \\ \leq\;& |\eta|\, C_6(C_7+C_8)t^i \;+\; C_6\,\|\eta W_0\mathbf{e}_1 - P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1}\|\,.\end{aligned} \tag{6.12}$$

In case of part a) we have $W_0\mathbf{e}_1 = \mathbf{v}_1^R$ and $B_0\mathbf{b}_0(\alpha_0)^{-1} = M\mathbf{v}_1^R$ while $P_1 = I$, hence there exists $\eta \neq 0$ such that $\|\eta W_0\mathbf{e}_1 - P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1}\| = \|\eta\mathbf{v}_1^R - M\mathbf{v}_1^R\| = 0$ and (6.11) holds with $C_9 := |\eta|\, C_6(C_7+C_8)$.

For part b) we have $W_0\mathbf{e}_1 = P_2\mathbf{v}_1^R$ while again $B_0\mathbf{b}_0(\alpha_0)^{-1} = M\mathbf{v}_1^R$. Therefore, there exists $\eta \neq 0$ such that $\|\eta W_0\mathbf{e}_1 - P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1}\| = \|\eta P_2\mathbf{v}_1^R - P_1^{-1}M\mathbf{v}_1^R\| = 0$ and (6.11) holds with $C_9 := |\eta|\, C_6(C_7+C_8)$.

Now for part c) we have $W_0\mathbf{e}_1 = P_2\mathbf{v}_1^R$ and also $B_0\mathbf{b}_0(\alpha_0)^{-1} = P_2\mathbf{v}_1^R$. We set $\eta = 1$ and conclude that $\|W_0\mathbf{e}_1 - P_1^{-1}B_0\mathbf{b}_0(\alpha_0)^{-1}\| = 0$, as a result (6.11) holds with $C_9 := C_6(C_7 + C_8)$.

To prove part d) for GY we use that $\mathbf{r}_{GY}^i = -\varphi^i\mathbf{r}^i$ and that $|\varphi^i| \leq const$, hence (6.11) holds with $C_9 = const C_6 \|P_1^{-1}\| \mathcal{R}^*$. □

We point out that in case of A and M large, Theorem 6.3 and Lemma 6.8 are of practical use only if $|\Gamma| \ll n$. Hence the following comments assume that $|\Gamma| \ll n$, possibly $\Gamma = \{1\}$.

Remark 6.9 *If* $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \leq C_9 t^i$ *then* $\mathcal{L}^i \propto \log(t^i/t^{i+1})$, *so the number of inner-iteration is related to the improvement the linear solve provides for the outer method. Further in this case we expect that for a linear converging method the number of inner-iterations is constant.*

We made this remark merely in contrast to the following one, which deals with the case $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \not\to 0$, which we did not consider in Lemma 6.8.

Remark 6.10 *If* $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \not\to 0$ *then the number of inner-iterations,* $\mathcal{L}^i$, *increases with the progress of the outer method, and* $\mathcal{L}^i \propto \log(1/t^{i+1})$.

As $\mathcal{L}^i$ is independent of t^i, it does not matter how good the current approximation is, the cost of the next solve depends only on the final tangent. We will illustrate this, for example, in Test 6.1.

Remark 6.11 *The condition* $\angle(\mathbf{v}_1^R, M\mathbf{v}_1^R) = 0$ *as in Lemma 6.8 part a) is always satisfied for the standard eigenvalue problem. However, if the GEP is derived by a FEM discretisation the 'mass' matrix M is a discretisation of the identity operator, hence* $\angle(\mathbf{v}_1^R, M\mathbf{v}_1^R) = 0$ *might be satisfied for some eigenvalue problems.*

Remark 6.12 *Part b) of Lemma 6.8 might be used to produce an optimal preconditioner for inexact inverse iteration using GMRES. If* $\angle(P_2\mathbf{v}_1^R, P_1^{-1}M\mathbf{v}_1^R) = 0$ *while* $P_1^{-1}(A - \mu M)P_2^{-1} = I + E$ *with* $\|E\|$ *small and* $|\lambda_1 - \mu|$ *small then we expect, for example, RQId using GMRES to perform exceptionally well.*

We did not comment on the case where $P_1^{-1}B_0\mathbf{b}_0$ lies in a small invariant subspace (invariant with respect to $P^{-1}(A - \lambda_1 M)P_2^{-1}$), but not in span$(W_0\mathbf{e}_1)$, as we regard this case as non-practical.

6.4 Overall costs

So far we discussed the cost per outer iteration which is essential to understand the differences between the methods. In order to discuss the overall efficiency we now present our second key result regarding the efficiency of inexact inverse iteration. Here we present bounds for the overall cost $\mathcal{T}$ as defined in Definition 6.2.

Theorem 6.13 *Let the conditions of Theorem 6.3 be satisfied. Further let $\mathcal{N}$ be the number of outer iterations used by the applied method to achieve the tangent t^* where $t^* < t^0$.*

a) If $\angle(M\mathbf{v}_1^R, \mathbf{v}_1^R) = 0$ then for the methods InvitFd, RQIf, RQId, InvitWf and InvitWd using unpreconditioned GMRES there exists $C_9 > 0$ with

$$\mathcal{T} \leq \mathcal{N}\left(1+|\Gamma| + \frac{\log(C_5 C_9 C_{10})}{\log(1/q_\Gamma)}\right) + \frac{\log(t^0/t^*)}{\log(1/q_\Gamma)}. \tag{6.13}$$

b) If $\angle(P_2\mathbf{v}_1^R, P_1^{-1}M\mathbf{v}_1^R) = 0$ then for the methods InvitFd, RQIf, RQId, InvitWf and InvitWd using preconditioned GMRES there exists $C_9 > 0$ such that (6.13) holds.

c) For the methods PInvit, GY there exists $C_9 > 0$ such that (6.13) holds.

d) For all other cases there exists $C_{11} > 0$ with $\|Q_\Gamma(W^i)^{-1}P_1^{-1}B^i\mathbf{b}^i(\alpha^i)^{-1}\| \leq C_{11}$ and hence

$$\mathcal{T} \leq \mathcal{N}\left(1+|\Gamma| + \frac{\log(C_5 C_{10} C_{11})}{\log(1/q_\Gamma)}\right) + \sum_{i=0}^{\mathcal{N}-1} \frac{\log(1/t^{i-1})}{\log(1/q_\Gamma)}. \tag{6.14}$$

Proof: Combining Theorem 6.3 with Lemma 6.8 using $\sum \log(t^i/t^{i+1}) = \log(t^0/t^*)$ proofs parts a) to c). Part d) follows immediately from Theorem 6.3. □

Again, for A, M large the result is only of interest if $|\Gamma| \ll n$. Therefore we assume that $|\Gamma| \ll n$.

Obviously, $\mathcal{T} \propto \log(t^0/t^*)$ is better than $\mathcal{T} \propto \sum \log(1/t^{i+1})$ is. However for all cases in Theorem 6.13 the major cost term is linear in the number of outer iterations $\mathcal{N}$, therefore it is sensible to use a method which reaches t^* with a small number of outer iterations, $\mathcal{N}$. Further reducing the number of outer iterations, reduces the number of terms in $\sum \log(1/t^{i+1})$.

Remark 6.14 *To reduce $\mathcal{T}$ it is vital to reduce the number of outer-iterations, as long as the number of inner-iterations per outer-iteration does not exceed limitations, for example, posed by available storage.*

In Chapter 3 we showed for the standard symmetric eigenvalue problem, that $\mathcal{N}$ is at least linear in $\log(t^0/t^*)$ and superlinear if the (outer) method is of higher order. The same is true for the GEP and the proof is similar to the one of Lemma 3.2. As a result we expect methods of higher order to be more efficient than linearly converging methods.

So far we did not comment on the methods using the two sided approach GRQIf and PInvitGRQ. In both methods two linear systems need to be solved per iteration,

one for the right eigenvalue problem and one for the left eigenvalue problem. We can apply our definitions and results directly to both methods and use subindices L and R to denote corresponding variables. Then the overall cost is given by $\mathcal{T}_L + \mathcal{T}_R$ where $\mathcal{T}_L = \sum \mathcal{L}_L^i$ and $\mathcal{T}_R = \sum \mathcal{L}_R^i$. As most of the bounds remain unchanged for the left problem we have the following result.

Remark 6.15 *For methods based on the generalised Rayleigh quotient we obtain under similar conditions as in Theorem 6.13 for the left and right eigenvalue problem the following bounds.*

a) For PInvitGRQ there exist constants C_5, C_9 and C_{10} such that

$$\mathcal{T}_{L+R} \leq 2\mathcal{N}\left(1+|\Gamma|+\frac{\log(C_5C_9C_{10})}{\log(1/q_\Gamma)}\right)+\frac{\log(t_L^0/t_L^{\mathcal{N}})+\log(t_R^0/t_R^{\mathcal{N}})}{\log(1/q_\Gamma)}, \quad (6.15)$$

b) For GRQIf and GRQId there exist constants C_5, C_9 and C_{10} such that

$$\mathcal{T}_{L+R} \leq 2\mathcal{N}\left(1+|\Gamma|+\frac{\log(C_5C_9C_{10})}{\log(1/q_\Gamma)}\right)+\sum_{i=0}^{\mathcal{N}}\frac{\log(1/t_L^{i+1})+\log(1/t_R^{i+1})}{\log(1/q_\Gamma)}, \quad (6.16)$$

Remark 6.16 *We expect the methods GRQIf and PInvitGRQ to be about twice as expensive per outer iteration as RQId.*

So far we did not discuss the efficiency of ICMf. However, see Remark 4.12, ICMf is a special case of GY.

Remark 6.17 *The efficiency result for GY applies also to ICMf.*

Remark 6.18 *If A is nonsymmetric then of all the methods discussed here using the standard approach, RQId and InvitWd are the most efficient methods. Further we expect RQId and InvitWd to be even more efficient than the other methods we discussed.*

Remark 6.19 *In case A is symmetric then the RQ is quadratic in t^i and hence RQIf, RQId and PInvit converge at least quadratic. Therefor we expect PInvit to be most efficient.*

In the symmetric case it is sensible to replace GMRES by MINRES as linear solver.

6.5 Tests

In order to support our theory we consider three examples. The first is the small constructed eigenvalue problem we used in Chapter 4. This problem will be used to show the behaviour of $\mathcal{L}^i$ and $\mathcal{T}$ for each method. However, the size of the problem restricts its use for a comparison between the methods in sense of which method is

more efficient for large problems. The second and third example are standard eigenvalue problems using medium sized matrices from matrix market. Both examples give reasonable insight in the overall efficiency and allow the comparison of the discussed methods.

Before we discuss the tests we recall briefly the definition of the methods made in Chapter 4.

6.5.1 Definition of Methods

Invit stands for inverse iteration with fixed shift and decreasing tolerance, that is Algorithm 5 with $\sigma^i = \varrho^0$ and $\tau^i = \min\{\widetilde{C}_3\ |\varrho^i|^{-1}\|r^i\|_2, \tau_0\}$.

RQIf is the Rayleigh quotient iteration with fixed tolerance, that is Algorithm 5 with $\sigma^i = \varrho^i$ and $\tau^i = \tau_0$.

RQId is the Rayleigh quotient iteration with decreasing tolerance, that is Algorithm 5 with $\sigma^i = \varrho^i$ and $\tau^i = \min\{\widetilde{C}_3\ |\varrho^i|^{-1}\|r^i\|_2, \tau_0\}$.

InvitWf is inexact inverse iteration using the Wilkinson update with fixed tolerance, that is Algorithm 6 with $\sigma^{i+1} = (\mathbf{z}^H A\mathbf{y}^i)/(\mathbf{z}^H M\mathbf{y}^i)$ and $\tau^i = \tau_0$.

InvitWd is inexact inverse iteration using the Wilkinson update with decreasing tolerance, that is Algorithm 6 with $\sigma^{i+1} = (\mathbf{z}^H A\mathbf{y}^i)/(\mathbf{z}^H M\mathbf{y}^i)$ and $\tau^i = \min\{\widetilde{C}_3\ |\ \varrho^i|^{-1}\|r^i\|_2, \tau_0\}$.

PInvit uses the modified right-hand side, $\mathbf{b}^i = P\mathbf{x}^i$, so Algorithm 5 with $\sigma^i = \varrho^i$ and $\tau^i = \tau^0$.

ICMf is the Inverse Correction Method with fixed shift, the first iteration is inexact inverse iteration, Algorithm 7 with $\sigma^0 = \varrho^0$, then inverse correction, Algorithm 7 with $\sigma^i = \varrho^0$ and $\tau^i = \tau_0$.

GY uses one iteration of InvitFd and then Algorithm 8 with $\tau^i = \min\{\tau_0, \widetilde{C}_3\ \|\mathbf{r}^i\|\}$, and $\sigma^i = \varrho^0$.

GRQIf is RQIf simultaneously applied to the left and the right eigenvalue problem and shift equaling the GRQ, that is Algorithm 9 with $\overline{\sigma_L^i} = \sigma_R^i = \varrho_G^i$, and $\tau_L^i = \tau_R^i = \tau_0$.

PInvitGRQ uses a modified right-hand side for the left and the right system, so Algorithm 9 with $\overline{\sigma_L^i} = \sigma_R^i = \varrho_G^i$, and $\tau_L^i = \tau_R^i = \tau_0$, while $\mathbf{b}_L^i = P_L\mathbf{x}_L^i$ and $\mathbf{b}_R^i = P_R\mathbf{x}_R^i$.

6.5.2 A small example

We use again our example from Chapter 4, for the full description see Section 4.5.1.

Hobbit The matrices A and M are real 62×62 matrices with $A = VDV^{-1}$ and $M = \text{diag}(1.1, 1.2, 1.3, \ldots, 7.2)$ where $V = U + 0.2 * I$ with U a full matrix of uniformly in $(0, 1)$ distributed random variables and $D = diag(D_1, D_2, \ldots, D_6, 1, 2, 3, \ldots, 50)$. Further the matrices D_j are 2×2 real matrices corresponding to the complex eigenvalues of A, $1 \pm 5i$,$1 \pm 1i$, $3 \pm 3i$, $3 \pm 1i$, $5 \pm 5i$ and $5 \pm 1i$. The non-standard construction of V is used to keep a moderate conditioning of the eigenvectors. Figure 4.5.1 shows a plot of the spectrum of the matrix pair A, M, the red asterisks indicate the four eigenvalues of interest.

Here we restrict attention to the two complex eigenvalues, $-0.77 + 4.09i$ and $2.32 + 0.43i$, to which we refer to as the extreme (complex) and the interior (complex) eigenvalue.

We point out that $\angle(\mathbf{v}_1^R, M\mathbf{v}_1^R) \neq 0$ and also $\angle(P_2\mathbf{v}_1^R, P_1^{-1}M\mathbf{v}_1^R) \neq 0$. Here we use this example merely to illustrate the behaviour of $\mathcal{L}^i$.

Test 6.1 *We apply InvitFd to the example 'Hobbit' to calculate the extreme and the interior eigenvalue. For the extreme eigenvalue we use unpreconditioned GMRES and for the interior eigenvalue left-preconditioned GMRES. In Table 6.2 we present results from two test runs each, using different choices for* $\widetilde{C}_3$ *in* $\tau^i \leq \min(\widetilde{C}_3(\varrho^i)^{-1} \|\mathbf{r}^i\|, \tau_0)$.

We observe from Table 6.2 that for all test runs the number of inner-iterations $\mathcal{L}^i$ increases with i. This increase in $\mathcal{L}^i$ is expected as $\angle(P_2\mathbf{v}_1^R, P_1^{-1}M\mathbf{v}_1^R) \neq 0$, and hence Remark 6.10 applies. Corresponding to the Theorem 6.3, the expected increase is $\mathcal{L}^{i+1} - \mathcal{L}^i \propto \log(t^i/t^{i+1})/\log(q_r)$ and is independent of the choice for $\widetilde{C}_3$. Further we observe that the increase in $\mathcal{L}^i$ slows down with $i \to \infty$, so, for example, in the left column of Table 6.2 the increase of 5 in the early stages reduces to an increase of 3 in the later part. This effect is due to the size of our example and the discrete nature of the spectrum of the preconditioned iteration matrix. As this effect is independent of $\widetilde{C}_3$ we expect that the slow down of the increase effects both runs in the same way, which is confirmed by the data in Table 6.2. For large problems we might see such an effect in a transitional early part but expect that the rate of increase converges to a positive constant.

In Remark 6.4 we explained that reducing C_3 in Theorem 4.2 improves the outer convergence but leads to an increase in $\mathcal{L}^i$. As $\widetilde{C}_3 \approx C_3const$ we are not surprised to observe this behaviour in Table 6.2, when for example comparing column three and four.

In Remark 6.14 we stated that reducing the number of outer-iterations $\mathcal{N}$ is vital in order to reduce the total number of inner-iterations $\mathcal{T}$. Comparing the difference in $\mathcal{T}$

	complex extreme				complex interior			
$\widetilde{C}_3$	2		0.2		2		0.2	
i	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$
0	5.0e-02	9	5.0e-02	13	5.0e-02	28	5.0e-02	29
1	1.7e-01	14	5.6e-02	23	1.4e-02	33	1.1e-02	37
2	2.7e-02	20	2.5e-03	28	2.7e-03	36	1.5e-03	39
3	1.9e-03	25	1.4e-04	33	3.9e-04	38	2.4e-04	41
4	1.3e-04	29	8.9e-06	38	6.6e-05	40	3.7e-05	43
5	1.9e-05	33	5.3e-07	42	1.2e-05	42	6.0e-06	44
6	2.0e-06	37	3.3e-08	46	2.2e-06	43	9.8e-07	46
7	1.4e-07	41	2.1e-09	49	4.2e-07	45	1.6e-07	47
8	9.6e-09	44	1.3e-10	52	7.8e-08	47	2.7e-08	49*
9	1.0e-09	47	8.2e-12	55	1.4e-08	48	4.7e-09*	50*
10	5.8e-11	50	5.1e-13	62	2.5e-09	49*	8.2e-10*	51*
11	5.5e-12	53	3.1e-14		4.6e-10*	50*	1.4e-10*	53
12	3.2e-13	56			9.5e-11*	51*	2.5e-11	54
13	1.9e-14				1.8e-11*	53	4.5e-12	60
14					3.3e-12	54	7.8e-13	
15					5.8e-13	60		
16					9.1e-14			
$\mathcal{T}$		458		441		717		643

Table 6.2: InvitFd applied to 'Hobbit' (Test 6.1)

for the two runs for the extreme complex eigenvalue we see that reducing $\mathcal{N}$ pays off. However the effect on $\mathcal{T}$ is minor in this case as the eigenvalue is an extreme eigenvalue which is well separated from the remaining ones. As a result of this the constants C_5, C_9 and C_{10} are small while $\Gamma = \{1\}$ is a good choice (and valid with respect to restrictions on Γ). Hence the cost term linear in $\mathcal{N}$ is small. In case of a less nicely separated eigenvalue, or interior eigenvalue and for larger examples (almost) in general we expect the difference to be larger.

The difference in $\mathcal{T}$ gets more apparent when we compare RQIf and RQId with InvitFd or later when we consider larger examples in Section 6.5.3.

Test 6.2 *We now apply RQIf and RQId as well as InvitWf and WinvitWd to the interior complex eigenvalue of 'Hobbit'. we present the corresponding results in Table 6.3.*

We explained in Chapter 4 (p. 96) why we expect that RQIf needs fewer outer iterations, $\mathcal{N}$, than InvitFd does. Here we observe that this difference pays off with respect to the overall efficiency. The test run with the least total number of inner iterations for InvitFd, had $\mathcal{T} = 615$ in case of the complex interior eigenvalue problem, we omit the data. In our results for RQIf and RQId we observed considerably smaller values for $\mathcal{T}$.

	RQIf		RQId		InvitWf		InvitWd	
	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$
0	5.0e-02	32	5.0e-02	28	5.0e-02	30	5.0e-02	32
1	5.2e-03	38	1.4e-02	39	7.8e-03	37	5.2e-03	41
2	3.8e-05	42	9.2e-05	49*	1.3e-04	41	3.3e-05	51*
3	5.2e-07	46	4.6e-09*	56	2.2e-06	45	1.1e-09*	56
4	1.2e-08	50*	2.7e-14		3.0e-08	49*	3.5e-14	
5	3.9e-11*	53			1.3e-10*	52		
6	4.4e-13	56			1.9e-12	55		
7	2.1e-14				2.4e-14	56		
8					2.08e-14			
$\mathcal{T}$		317		172		365		180

Table 6.3: RQIf, RQId, InvitWf and InvitWd applied to 'Hobbit', (Test 6.2)

Also from Table 6.3 we observe that the difference with respect to $\mathcal{L}^i$ and $\mathcal{T}$ between using the RQ as shift and using the Wilkinson update as shift is insignificant.

We can compare the results in Table 6.3 with the right two columns in Table 6.2. Most importantly we observe that $\mathcal{L}^i \propto \log(1/t^{i+1})$ as stated in Remark 6.10. Secondly we observe that for any fixed t^{i+1}, the value of $\mathcal{L}^i$ is almost independent of the applied method. So all methods using the standard right-hand side show the same behaviour for $\mathcal{L}^i$. Finally by comparing those data marked with asterisks, all showing $49 \leq \mathcal{L}^i \leq 51$, we observe that t^{i+1} is least favourable for the quadratically converging methods, RQId and InvitWd.

The effects we reported about in Test 6.1 were also observed for RQIf, RQId, InvitWf and InvitWd, however we omit the corresponding results.

Test 6.3 *We repeat Test 6.2 for PInvit, ICMf and GY, for the results see Table 6.4.*

In Test 4.3 we already observed the poor rate of convergence for PInvit. Here, the more important observation is that $\mathcal{L}^i$ is constant for all three methods, which is as predicted in Remark 6.9. We observe that tightening the residual constraint for PInvit does not improve the outer convergence but leads to more inner iterations. As the rate of convergence for both test runs is about the same as for exact solves hence solving the system more accurately increases the cost per solve without improving the convergence. The results presented here are not sufficient to support the claim that $\mathcal{L}^i \propto \log(t^i/t^{i+1})$.

The same can be reported for GY and ICMf, see Table 6.4. In Chapter 4 we reported that the outer convergence of GY and ICMf is similar and related to the convergence for InvitFd. First we observe that the difference between ICMf and GY with respect to the efficiency is negligible. Comparing ICMf and GY with InvitFd with respect to the number of inner iterations, the advantage of solving the correction equations is apparent. As a result ICMf and GY are significantly more efficient than InvitFd is. As

	PInvit $\tau_0 = 0.2$		PInvit $\tau_0 = 0.05$		ICMf		GY	
i	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$	t^i	$\mathcal{L}^i$
0	5.0e-02	28	5.0e-02	32	5.0e-02	32	5.0e-02	32
1	1.4e-02	28	5.2e-03	31	5.2e-03	30	5.2e-03	31
2	3.1e-03	27	1.2e-03	29	1.5e-03	33	2.4e-03	33
3	6.6e-04	24	3.8e-04	29	2.6e-04	32	4.0e-04	33
4	2.6e-04	25	1.1e-04	29	4.8e-05	32	6.7e-05	30
5	5.4e-05	24	3.3e-05	28	9.2e-06	32	1.5e-05	33
6	2.1e-05	26	9.7e-06	28	1.6e-06	32	2.6e-06	33
7	6.3e-06	27	2.7e-06	28	2.9e-07	32	4.7e-07	32
8	1.1e-06	24	7.9e-07	28	5.7e-08	32	8.5e-08	32
9	4.8e-07	24	2.3e-07	28	1.0e-08	32	1.7e-08	32
10	1.2e-07	27	6.8e-08	28	1.9e-09	32	3.3e-09	32
11	2.3e-08	24	2.0e-08	28	3.7e-10	32	6.0e-10	31
12	1.1e-08	25	5.8e-09	28	6.9e-11	32	1.1e-10	32
13	3.4e-09	27	1.7e-09	28	1.2e-11	32	2.1e-11	32
14	6.8e-10	24	5.0e-10	28	2.4e-12	32	3.9e-12	32
15	2.5e-10	25	1.4e-10	28	4.5e-13	32	7.1e-13	31
16	5.0e-11	24	4.3e-11	28	7.0e-14	32	1.3e-13	32
17	2.6e-11	27	1.2e-11	28	1.2e-14		2.9e-14	
18	5.0e-12	24	3.7e-12	29				
19	1.8e-12	27	1.1e-12	28				
20	3.6e-13	23	3.4e-13	28				
21	1.9e-13	25	9.8e-14	28				
22	4.9e-14	26	3.5e-14	24				
23	1.8e-14		1.9e-14					
$\mathcal{T}$		585		651		543		543

Table 6.4: PInvit, ICMf and GY applied to 'Hobbit', (Test 6.3)

	GRQIf $\tau^0 = 0.01$			GRQIf $\tau^0 = 0.001$			PInvitGRQ $\tau^0 = 0.2$		
i	t^i	$\mathcal{L}_R^i$	$\mathcal{L}^i$	t^i	$\mathcal{L}_R^i$	$\mathcal{L}^i$	t^i	$\mathcal{L}_R^i$	$\mathcal{L}^i$
0	5.0e-02	34	70	5.0e-02	37	76	5.0e-02	24	56
1	3.2e-03	42	84	3.0e-03	44	88	2.7e-02	31	64
2	1.4e-06	53	107	1.2e-06	54	109	1.3e-03	37	75
3	4.1e-13	55	110	1.1e-13			5.3e-07	45	90
4	1.6e-13						9.0e-14		
$\mathcal{T}$			371			273			285

Table 6.5: GRQIf and PInvitGRQ applied to 'Hobbit', (Test 6.4)

RQIf has the same behaviour for $\mathcal{L}^i$ as InvitFd has, it is the reduced number of outer iterations which makes it more efficient than ICMf and GY.

We now compare the test runs of PInvit, ICMf and GY with the results for InvitFd applied to the interior eigenvalue, so Table 6.3 with right two columns of Table 6.3. In the first two outer-iterations, InvitFd needs about the same number of inner-iterations, which is not surprising as $1/t^{i+1}$ is moderate. While the number of inner iteration increases to more than double the initial value for InvitFd, the number of inner iterations remains constant for PInvit, ICMf and GY. As a result ICMf and GY which exhibit the same rate of convergence as InvitFd, are significantly more efficient than InvitFd. Also PInvit is more efficient than InvitFd despite needing 6 or 9 outer-iterations more.

Test 6.4 *We repeat Test 6.2 using GRQIf and PInvitGRQ, the corresponding results are given in Table 6.5.*

In Table 6.5 where we reported the results for the methods GRQIf and PInvitGRQ. We tabulated in addition to the number of inner-iterations for the *right* solve $\mathcal{L}_R^i$ also the combined number of inner-iterations $\mathcal{L}^i = \mathcal{L}_L^i + \mathcal{L}_R^i$. As for all methods we run GRQIf with different values for τ^0. Here for GRQIf we present the best and the worst result we experienced. In case of $\tau^0 = 0.001$ we obtained an acceptable approximation in three outer iterations, which leads to $\mathcal{T} = 273$. However choosing $\tau^0 = 0.2$, the result is omitted here, we obtain in four outer-iterations a satisfactory solution with $\mathcal{T} = 342$. The worst result was obtained for $\tau^0 = 0.01$ with $\mathcal{T} = 371$. With decreasing τ^0 we obtained higher costs $\mathcal{T}$ until we simultaneously observed a reduction in $\mathcal{N}$ from 4 to 3. This links nicely with the theory, as Theorem 6.13 states that $\mathcal{T} \propto \sum \log(1/t^{i+1})$ as long as $\mathcal{N}$ is fixed. Now reducing τ^0 leads to a better convergence and therefore reduces, for example, $t^{\mathcal{N}-1}$ and thereby increases $\mathcal{T}$. Hence if reducing τ^0 leads to a smaller number of outer-iterations then $\mathcal{T}$ is reduced otherwise $\mathcal{T}$ increases. This effect is not limited to GRQIf, but for the presented data, it is the most prominent example. In our tests we experience this effect for all methods using the standard right-hand

side.

In Table 6.5, we observe for PInvitGRQ, that the number of inner-iterations for $\mathcal{L}^i$ is about twice the for $\mathcal{L}_R^i$, except for the first iteration. As the angle between the left and right eigenvector is almost $\pi/2$, the initial vector $\mathbf{x}_L^0 = \mathbf{x}_R^0$ was a good approximation for the right eigenvector but not for the left eigenvector. As the same shift is used for the left and the right equation we expect $t_L^i \approx t_R^i$ for $i \geq 1$. Now $\mathcal{L}_L^i \propto \log(t_L^i/t_L^{i+1}$ and $\mathcal{L}_R^i \propto \log(t_R^i/t_R^{i+1}$ according to Remark 6.9 we expect that besides the first iteration $\mathcal{L}_L^i \approx \mathcal{L}_R^i$.

6.5.3 Two further tests

The previous example does not necessarily give a good indication about the overall efficiency for large problems. Hence we consider now two medium sized examples to support our Remarks on the overall cost.

Olmstead The matrix A is the unsymmetric 1000×1000 matrix 'olm1000' from 'Matrix Market'[1] while the matrix M is the identity matrix. We consider here four eigenvalues of this matrix, two real and two complex. In Figure 6-1 we provide a plot of the right most eigenvalues of A, the eigenvalues of interest are indicated by red asterisks. Not included in the plot are 484 real eigenvalues lying in the interval $(-10164, -15)$. The eigenvalues of interest are tabled below, together with the corresponding gaps.

	extreme real	interior real	extreme complex	interior complex
λ_1	4.5	0.09	1.3+2i	-0.35+4.7i
gap	0.62	0.32	1.2	0.99

Tolosa The matrix A is the 1090×1090 matrix 'tols1090' from 'Matrix Market'[1] while the matrix M is the identity matrix. In Figure 6-2 we provide two plots of the eigenvalues of A, the left plot giving the complete spectrum and the right plot all eigenvalues with real part larger -10, the eigenvalues of interest are indicated by red asterisks. The two complex eigenvalues of interest are tabled below, together with the corresponding gaps. For both eigenvalues we consider a preconditioner designed for the 'extreme' eigenvalue, that is $P \approx= A - 150iI$.

	extreme complex	interior complex
λ_1	-0.15+156i	-0.25+26.5i
gap	6	7.2

Test 6.5 *We now apply all methods to the four eigenvalues of interest of the example 'Olmstead'. For this test we only provide the resulting values for $\mathcal{N}$ and $\mathcal{T}$ in Table 6.6.*

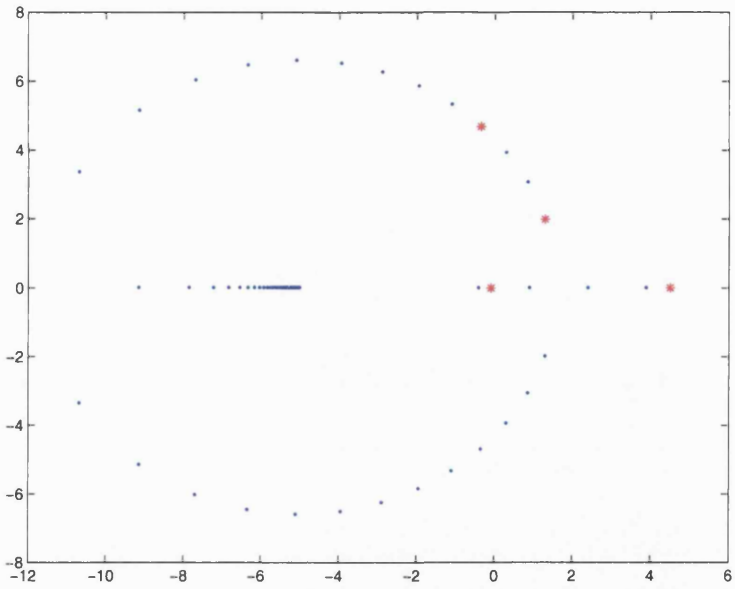


Figure 6-1: Right most eigenvalues of A for example 'Olmstead', red asterisks indicate eigenvalues of interest

	extreme real		interior real		extreme complex		interior complex	
	$(\mathcal{N})$	$\mathcal{T}$	$(\mathcal{N})$	$\mathcal{T}$	$(\mathcal{N})$	$\mathcal{T}$	$(\mathcal{N})$	$\mathcal{T}$
InvitFd	(30)	1535	(9)	587	(14)	964	(22)	1835
RQIf	(3)	124	(3)	187	(4)	244	(4)	289
RQId	(3)	115	(3)	158	(3)	159	(3)	201
InvitWf	(4)	178	(4)	224	(4)	240	(5)	346
InvitWd	(3)	129	(3)	192	(3)	177	(3)	220
PInvit	(6)	198	(-)	-[1]	(-)	-[2]	(-)	-[2]
GY	(35)	1029	(-)	-[3]	(-)	-[3]	(30)	1810
ICMf	(30)	500	(20)	1083	(20)	922	(30)	1622
GRQIf	(3)	241	(3)	331	(3)	321	(3)	421
PInvitGRQ	(3)	233	(4)	321	(5)	409	(5)	543

Table 6.6: Total number of inner iterations $\mathcal{T}$ and number of outer iterations $(\mathcal{N})$ for example 'Olmstead', (Test 6.5)

In Table 6.6 we report only one test run for each method, however we carried out several tests using different starting vectors and different choices for the parameters. There were no significant differences between different test runs.

One key observation to be made is that RQId and InvitWd are according to the results presented in Table 6.6 the most efficient methods.

Comparing RQIf with InvitWf and RQId with InvitWd we observe than the methods using the RQ perform better that those using the Wilkinson update. An exception to this is the complex extreme eigenvalue where InvitWf is more efficient than RQIf. We observe that for the two real eigenvalues RQIf outperforms InvitWd, however this is due to the fact that these eigenvalues are well conditioned, and hence RQIf exhibits in these cases superlinear convergence .

The encouraging results for PInvit in the case of the extreme real eigenvalue are overshadowed by the convergence problems we experience for the other three cases. For the real interior eigenvalue, marked by [1], we obtained no convergence. In case of the two complex eigenvalues, marked by [2], the convergence was so slow that we stopped after 70 iterations without having converged to the specified tolerance, however convergence was eventually achieved. While the above discussed problems are convergence problems with the outer method, we encountered a lack of convergence of GMRES for the extreme complex eigenvalue, due to a very tight inner tolerance.

From Table 6.6 we observe that InvitFd, GY and ICMf perform poorly. Their high values for $\mathcal{T}$ are unsurprising as $\mathcal{T}$ is linear in $\mathcal{N}$ which is large itself. So the poor outer convergence of theses three methods makes them inefficient, which is what we expect according to Remark 6.14.

Unsurprisingly we observe that the total number of inner-iterations for GRQIf, so including left and right solves, is about double the value as for RQId. The initially poorer outer convergence of PInvitGRQ results in high values for $\mathcal{T}$.

Test 6.6 *We repeat Test 6.5 for the example 'Tolosa', the resulting values for $\mathcal{N}$ and $\mathcal{T}$ are given in Table 6.6.*

The preconditioner was constructed with respect to the extreme eigenvalue, $P = A + 150iI + E$ where $\|E\|_2 = 0.6$. This relative accurate preconditioner was used with respect to PInvit, while the choice of shift for the preconditioner is important to keep the number of inner-iterations per outer-iteration small, not exceeding 30 iterations.

As a result for all tests the total number of inner-iteration $\mathcal{T}$ is low for all methods. applied to the extreme eigenvalue, see left column in Table 6.7. In contrast the total number of inner-iterations is significantly larger for the interior eigenvalue.

Again RQId and InvitWd are the most efficient methods. The higher cost of InvitWd compared with RQId for the interior eigenvalue is due to a better convergence as we now explain. Both methods start with the same tangent, their first tangent as

[1] http://gams.nist.gov/MatrixMarket

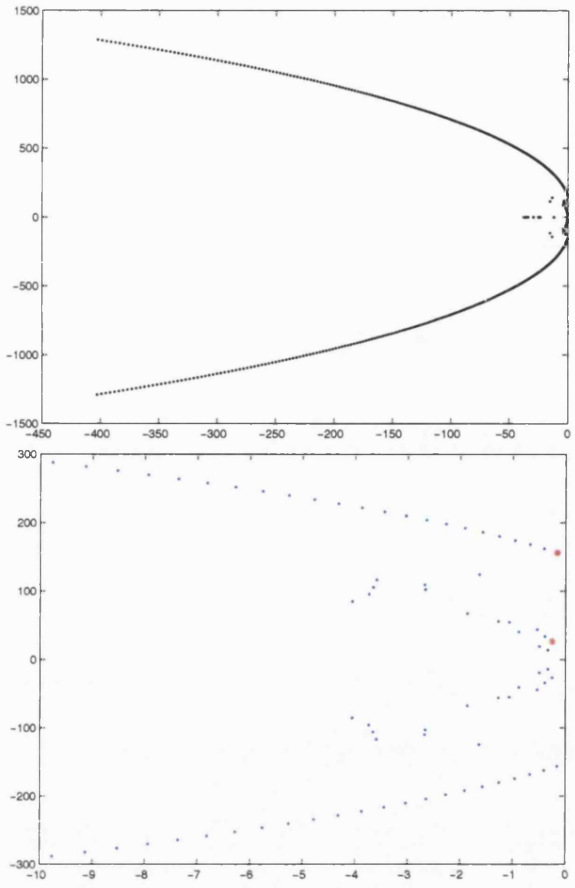


Figure 6-2: Eigenvalues of A for example 'Tolosa', red asterisks indicate eigenvalues of interest, upper plot full spectrum, lower plot eigenvalues with real part larger than -10

	extreme complex		interior complex	
	$(\mathcal{N})$	$\mathcal{T}$	$(\mathcal{N})$	$\mathcal{T}$
RQIf	(4)	51	(4)	255
RQId	(3)	41	(3)	163
InvitWf	(5)	55	(5)	288
InvitWd	(3)	47	(3)	185
PInvit	(6)	57	(5)	256
GY	(11)	91	(8)	443
ICMf	(11)	98	(7)	392
GRQIf	(3)	98	(3)	366
PInvitGRQ	(3)	94	(3)	333

Table 6.7: Total number of inner iterations $\mathcal{T}$ and number of outer iterations $(\mathcal{N})$ for example 'Tolosa', (Test 6.6)

well as their third (and final) tangent are of same order of magnitude. However, the tangent in iteration two, t^2, for InvitWd is more than two orders of magnitude smaller than t^2 for RQId. As the cost of InvitWd and RQId per solve $\mathcal{L}^i$ is proportional to $\log(1/t^{i+1})$ we would expect that the cost $\mathcal{L}^i$ is similar except for $\mathcal{L}^1$, which is what we observed in the test.

Here PInvit converges without difficulty and has a total number of inner-iterations similar to RQIf despite needing one or respectively two iterations more than RQIf. While all methods used only a few outer-iterations, still the need of a small number of outer-iterations is apparent.

6.6 Conclusion

Based on the convergence results in Chapter 4 and encouraged by the efficiency results in Chapter 3 we analysed the efficiency of inexact inverse iteration using GMRES being applied to the GEP. Our analysis is based on the assumption that the cost of a matrix vector product is significantly larger than the cost for orthonormalizing the GMRES basis vectors. Further we assumed that enough storage is available so that we can neglect any limitations in the number of inner-iterations. In practice these two assumptions are equal to requiring an excellent preconditioner, such that only a few iteration will be performed to solve the linear equations. Based on these assumptions we presented two key results concerning the efficiency.

The first result, Theorem 6.3, provides a bound on the number of GMRES iterations per outer-iteration. This result is general in the sense that it is valid for a large class of variations of inexact inverse iteration, only restricted by the conditions of Theorem 4.2, ensuring convergence, and equation (6.2). The a-posteriori bound provided by Theorem 6.3 links the cost of a linear solve with the progress achieved by the same solve. As a direct consequence of this link both sides contain a-priori unknown quantities, hence a-posteriori. However, it is this link between the cost of the linear solve and the progress the linear solve provides for the outer method which allows the theoretical study of the efficiency of inexact inverse iteration.

Based on this first result we made several remarks with respect to the behaviour of practical methods. So we showed that the cost of a linear solve for methods using the standard right-hand side in inexact inverse iteration, increases logarithmically with the achieved error angle for the eigenvalue problem. For methods using correction equations and our extension of the approach from Simoncini and Eldén (2002) using a modified right-hand side, we proved that the cost per solve depends only on the reduction of the error angle. Our theoretical observations with respect to the cost of a single linear solve illustrate the benefit of the cost being dependent on the reduction of the error angle rather than only being dependent on the achieved error angle.

Based on the first key result, Theorem 6.3, the second key result, Theorem 6.13, provides an upper-bound for the overall cost. This result is in contrast to the first only applicable to the methods Invit, RQIf, RQId, InvitWf, InvitWd, PInvit, and GY. However Remarks 6.15 and 6.17 extend it to PInvitGRQ, GRQIf and ICMf respectively. Like the first result, Theorem 6.13 states only an a-posteriori bound on the total cost. The importance of Theorem 6.13 is the fact that it leads to Remarks 6.14, 6.18 and 6.19, which state the main theoretical observations of the efficiency analysis. In order to extend Theorem 6.13 to further methods the results presented in Section 6.3 should prove to be useful.

Based on the second result we established that reducing the number of outer-iterations $\mathcal{N}$ is beneficial to reduce the overall number of inner-iterations $\mathcal{T}$, see Remark 6.14. Further we stated that for the unsymmetric eigenvalue problem the Rayleigh quotient iteration with decreasing tolerance, RQId, and inexact inverse iteration using the Wilkinson update and a decreasing tolerance, InvitWd, are the two most efficient methods of the methods considered here, see Remark 6.18. Both observations were confirmed by the numerical results. Further we stated that in case of the symmetric generalised eigenvalue problem we expect PInvit to be most efficient.

However, in this chapter we restricted ourselves to linear solves using either unpreconditioned GMRES or preconditioned GMRES. None of the methods considered here combines a minimal number of outer iteration with cost per inner-iterations only depending on the reduction of the error angle. Nevertheless we believe that such methods exists and reasonable candidates are, for example, RQId using augmented GMRES and alternating version of PInvit using GMRES and an alternating version of RQIf using augmented GMRES. By alternating version we mean a two sided approach with a shift update after each solve. So there is further research needed to incorporate augmented and deflated GMRES. Further this thesis should help the understanding of inexact inverse iteration with respect to convergence and with respect to efficiency.

In this Chapter, see Section 6.1, we assumed that the preconditioner is of such quality that the number of inner-iterations does not exceed restrictions posed by memory limitations nor that the cost of orthogonalizing the Krylov basis vectors in GMRES gets dominant. If these assumptions are not valid, our results still hold and show the kind of increase both in the number of outer-iterations and in the total number of inner-iterations one expects by limiting the number of inner-iterations per outer-iteration. To ease any restriction one might consider restarted GMRES or other solvers like QMR or BICGSTAB, however, for these methods the results presented here are not valid.

Our practical results show that the theoretical bounds are descriptive. For example, we observed that the number of inner-iterations per outer-iteration increases like $log(1/t^{i+1})$ for methods using the standard right-hand side $\mathbf{b}^i = M\mathbf{x}^i$, as we predicted in Remark 6.10. Further we were able to confirm Remark 6.9, stating that the cost

of a linear solve measured in the number of inner-iterations behaves like $log(t^i/t^{i+1})$ for PInvit, ICMf and GY. However, as we already pointed out in Chapter 4 the outer convergence of PInvit, ICMf and GY is not optimal. PInvit can suffer lack of convergence depending on the quality of the RQ and the quality of the preconditioner. Even in the case that both RQ and preconditioner are of a good quality we might observe a smaller area of convergence for PInvit than, for example, for RQId. The convergence of ICMf and GY is as expected only linear and hence too many outer iterations are needed. Further our practical results show the importance of a small number of outer-iterations thus supporting Remark 6.14. While GRQIf and PInvit converge with a small number of outer-iterations, which equals the one for RQId, they need twice as many inner-iterations as RQId and thus are not as efficient. However, if the left and right eigenvector are sought then PInvitGRQ is our first choice. Our results confirm that RQId is robust and (often) the most efficient method for the unsymmetric GEP.

To confirm that our theoretical observations and numerical results reflect the behaviour of inexact inverse iteration applied to large sparse eigenvalue problem further tests are needed. Also, to confirm Remark 6.19 we need tests with A symmetric and $M \neq I$ spd.

The effect of the preconditioner on the performance of inexact inverse iteration in general and PInvit in particular needs further exploitation.

A Practical Recommendation

The use of inexact inverse iteration as an eigenvalue solver in its own right might not be recommended. However, if the user can provide a good enough initial guess, perhaps by using the Arnoldi method, and the sought eigenvalue is known to be well separated then inexact inverse iteration might be a worthwhile alternative to the Jacobi-Davidson method and is able to outperform the Arnoldi method, see, for example, Graham et al. (2003).

If the eigenvalue problem $A\mathbf{x} = \lambda M\mathbf{x}$ is such that A is symmetric and M spd then we recommend the use of PInvit, see page 67 and page 100, with MINRES as linear solver. While we achieve good and stable performances with PInvit using a very relaxed stopping condition $\tau^0 = 0.8$ we recommend $\tau^0 = 0.1$ for stability reasons. We advise to build into MINRES additional stopping conditions to detect failure of the linear solver early on and to avoid unnecessary computations caused by solving the eigenvalue problem too accurately.

In case of the unsymmetric eigenvalue problem the situation is not as clear cut as for the symmetric. As PInvit is currently not reliable enough we recommend the use of RQId. If the conditioning of the eigenvalue is poor and thus the RQ not a good approximation of the sought eigenvalue we recommend the use of InvitWd instead. As in the symmetric case we advise that the GMRES algorithm be used with some

additional stopping conditions.

Appendix A

Chebyshev polynomials

A.1 Minimal polynomials

In this appendix we gather a few results on Chebyshev polynomials with respect to the min-max problem

$$\eta_k(D) := \min_{p \in \Pi_k^1} \max_{x \in D} |p(x)|, \tag{A.1}$$

where $D \subset \mathbb{C}$. The results presented here are taken from Chatelin (1993), Fischer and Peherstorfer (2001). Chebyshev polynomials are usually defined by the recurssion

$$\begin{aligned} T_0(z) &= 1, \\ T_1(z) &= z, \\ T_{k+1}(z) &= 2zT_k(z) - T_{k-1}(z). \end{aligned} \tag{A.2}$$

One can show that

$$T_k(z) = \cosh(k \cosh^{-1}(z)) \tag{A.3}$$

satiesfies the recursion (A.2).

In the remainder of this Appendix we provide bounds for the min-max problem (A.1), with respect to some specific domains. First we consider classical results for Chebyshev polynomials, for example see Chatelin (1993).

Theorem A.1 *Let $a, b, c \in \mathbb{R}$ with $ab > 0$ and $r, s \in \mathbb{R}^+$.*

1. *If $D = [a, b]$ (real line), set $r = \frac{1}{2}|b - a|$ and $c = \frac{1}{2}(a + b)$, then*

$$\eta_k(D) = \frac{1}{T_k(|c| / r)}, \tag{A.4}$$

2. *If* $D = \{z \in \mathbb{C} \mid |z-a| \leq r, |a| > r\}$ *(complex disc with real center), then*

$$\eta_k(D) = \left(\frac{r}{|a|}\right)^k, \tag{A.5}$$

3. *If* $D = \{z| \ |z-(c+s)| + |z-(c-s)| \leq 2r, |c| > r > s\}$ *(complex ellipse with real foci), then*

$$\eta_k(D) \ = \ \frac{T_k(r/s)}{T_k(|c| \ /s)}. \tag{A.6}$$

Proof: For the proof see Chatelin (1993, Theorem 6.6.2). □

To extend these results to more complicated domains one can use an idea as in Fischer and Peherstorfer (2001). Given $k \in \mathbb{N}$ and a set D then denote by $T_k^D(z)$ the solution of the Chebyshev approximation problem

$$\max_{z \in D} |T_k^D(z)| \ = \ \min_{f \in \Pi} \max_{z \in D} |z^n - f(z)| \, .$$

Then for all domains we have $\eta_k(D) \leq \max_{z \in D} |T_k^D(z)| \ / \ |T_k^D(0)|$. Further let φ be a polynomial of degree l with leading coefficient $a_l \neq 0$ and no multiple roots. Now Fischer and Peherstorfer (2001, Corollary 2.2) show that

$$T_k^{\varphi^{-1}(D)}(z) = a_l^{-k} T_{lk}^D(\varphi(z)). \tag{A.7}$$

So let $\varphi^{-1}(D)$ be a complex ellipse with real foci while D is some domain in $\mathbb{C}$ for which Theorem A.1 does not apply. Then (A.7) gives that $\eta_{kl}(D) \leq \eta_k(\varphi(D))$. In order to demonstrate the implication of this result we consider two different domains. First we take two real intervals, second a dumbell shaped domain. Given foci f_1 and $f_2 \in \mathbb{C}$ with $f_1 \neq f_2$ and radius $R \in \mathbb{R}^+$ then we define a dumbell as

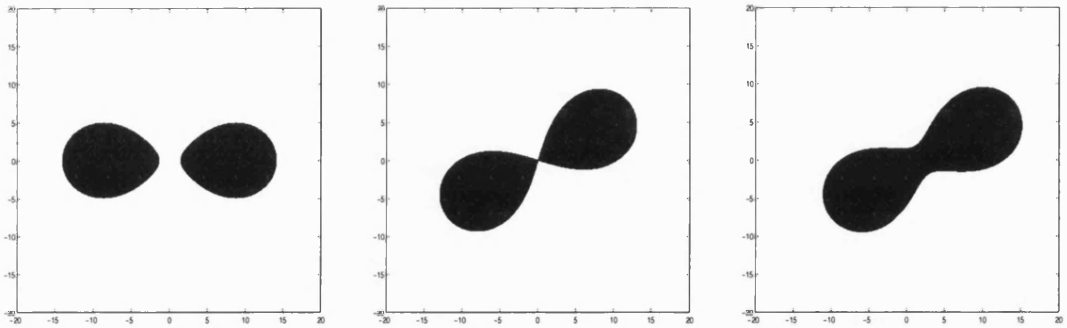
$$D := \{z| \ |z-f_1||z-f_2| \leq R^2\}. \tag{A.8}$$

Figure A-1 illustrates three dumbells with $R < R^*$, $R = R^*$ and $R > R^*$ where $R^* := \frac{1}{2} \ |f_1 - f_2|$. For $f_1 = f_2$ the dumbell equals a circle. Further a dumbell with $R > R^*$ includes the ellipse $\{z| \ |z-f_1| + |z-f_2| \leq 2r\}$ where $r \leq \sqrt{\frac{1}{2}R}$.

Corollary A.2 *Let* $a, b, c, d, R \in \mathbb{R}$ *and* $f_1, f_2 \in \mathbb{C}$.

1. *If* $D = [a,b] \cup [c,d]$ *with* $a < b < 0 < c < d$ *(two real lines), then set* $\alpha := \max\{1-(a-b)(a-c)(bc)^{-1}, 1-(d-b)(d-c)(bc)^{-1}\}$ *to obtain*

$$\eta_{2k}(D) \ \leq \ \frac{1}{T_k(|\frac{\alpha+1}{\alpha-1}|)}, \tag{A.9}$$



a) b) c)

Figure A-1: Dumbells, a) with $f_1 = -10$, $f_2 = 10$ and $R = 9.9$; b) with $f_1 = -8.8-4.8i$, $f_2 = 8.8+4.8i$ and $R = 10$; and c) with $f_1 = -8.8-4.8i$, $f_2 = 8.8+4.8i$ and $R = 10.2$

2. If $D = \{z|\ |z-f_1||z-f_2| \leq R^2\}$ *with* $|f_1 f_2| > R^2$ *then*

$$\eta_{2k}(D) \ \leq \ \frac{(R)^{2k}}{|f_1 f_2|^k}. \tag{A.10}$$

Proof: For part (1) consider $q(z) := 1-(z-b)(z-c)(bc)^{-1}$ then $q(0) = 0$ and $q(b) = q(c) = 1$ and as $bc < 0$, $q \to \infty$ for $z \to \ ^+_-\infty$. Then we can use Theorem A.1 part 1 with $c = \frac{1}{2}(\alpha+1)$ and $r = \frac{1}{2}(\alpha-1)$ together with (A.7) to obtain (A.9).

Now for (2) consider $q(z) = 1-(f_1-z)(f_2-z)(f_1 f_2)^{-1}$ then q(0) = 0 and $S := q(D) = \{z|\ |z-1| \leq \frac{R^2}{|f_1 f_2|}\}$. Using Theorem A.1 and (A.7) we obtain (A.10). □

A.2 Bounds on Chebyshev polynomials

So far we bounded $\eta_k(D)$ by Chebyshev polynomials, now we derive a bound for the Chebyshev polynomials. We then use this bounds for the sets discussed in Theorem A.1 and Corollary A.2.

As $|T_k(x)| = |T_k(-x)|$ for $x \in \mathbb{R}$ we assume in the following that $x > 0$. If $t = \cosh(x)$ then

$$T_k(t) \ = \ \cosh(kx) \ = \ \frac{1}{2}\left(e^{kx} + e^{-kx}\right).$$

For $kx > 0$ this leads to the bounds

$$\frac{1}{2}e^{kx} \ \leq \ T_k(t) \ \leq \ e^{kx}.$$

Next we use $e^x = \cosh(x) + \sinh(x) = t + \sqrt{t^2-1}$ to obtain $T_k(t) > \frac{1}{2}\left(t+\sqrt{t^2-1}\right)^k$,

and $T_k(t) < \left(t + \sqrt{t^2-1}\right)^k$. Now replace t by $(u+v)/(u-v)$ to get

$$\begin{aligned} T_k\left(\frac{u+v}{u-v}\right) &> \frac{1}{2}\left(\frac{\sqrt{u}+\sqrt{v}}{\sqrt{u}-\sqrt{v}}\right)^k \\ &= \frac{1}{2}\left(\frac{\sqrt{\frac{u}{v}}+1}{\sqrt{\frac{u}{v}}-1}\right)^k. \end{aligned} \tag{A.11}$$

Finally replace $u = \frac{1}{2}(x+y)$ and $v = \frac{1}{2}(x-y)$ which is the same as $x = u+v$ and $y = u-v$, giving

$$\left(\frac{\sqrt{\frac{x+y}{x-y}}+1}{\sqrt{\frac{x+y}{x-y}}-1}\right)^k > T_k\left(\frac{x}{y}\right) > \frac{1}{2}\left(\frac{\sqrt{\frac{x+y}{x-y}}+1}{\sqrt{\frac{x+y}{x-y}}-1}\right)^k. \tag{A.12}$$

Corollary A.3 *Let $a, b, c, d, r, s \in \mathbb{R}$ and $f_1, f_2 \in \mathbb{C}$.*

1. *For $D = [a,b]$ with $ab > 0$ let $\kappa := \max_{z\in D}|z| \,/\, \min_{z\in D}|z|$ then*

$$\eta_k(D) < 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k. \tag{A.13}$$

2. *For $D = [a,b] \cup [c,d]$ with $a < b < 0 < c < d$ let $\kappa := \max_{z\in D}|z| \,/\, \min_{z\in D}|z|$ then*

$$\eta_k \leq \frac{1}{2}\sqrt{\frac{\kappa+1}{\kappa-1}}\sqrt{\frac{\kappa-1}{\kappa+1}}^k. \tag{A.14}$$

3. *For $D = \{z \mid |z-(c+s)| + |z-(c-s)| \leq 2r\}$ with $|c| > r > |s| > 0$ let $\kappa_1 := \frac{r+|s|}{r-|s|}$ and $\kappa_2 := \frac{|c+s|+|c-s|+2|s|}{|c+s|+|c-s|-2|s|}$ then*

$$\eta_k \leq 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\frac{\sqrt{\kappa_1}+1}{\sqrt{\kappa_1}-1}\right)^k. \tag{A.15}$$

Proof: For part 1 we use (A.12) with part 1 of Theorem A.1 and for part A.14 use (A.11) with part 1 of theorem A.1. For part 3 we use Fischer and Freund (1990, Theorem2 and equation (1)) to obtain

$$\eta_k \leq \frac{T_k(\frac{r}{|s|})}{T_k(\frac{|c+s|+|c-s|}{2|s|})},$$

from which (A.15) follows using (A.12). □

For the case where D is an ellipse it is not obvious that $\eta_k \to 0$ for $k \to \infty$. Assume

$c > r > s > 0$ then

$$
\begin{aligned}
& r + s < c + s \\
\Leftrightarrow \quad & \tfrac{2s}{c+s} < \tfrac{2s}{r+s} \\
\Leftrightarrow \quad & \sqrt{1 - \tfrac{2s}{c+s}} > \sqrt{1 - \tfrac{2s}{r+s}} \\
\Leftrightarrow \quad & \sqrt{(c-s)(r+s)} > \sqrt{(r-s)(c+s)} \\
\Leftrightarrow \quad & \tfrac{\sqrt{(c+s)(r+s)} - \sqrt{(r-s)(c-s)} - \sqrt{(c-s)(r+s)} + \sqrt{(c+s)(r-s)}}{\sqrt{(c+s)(r+s)} - \sqrt{(r-s)(c-s)} + \sqrt{(c-s)(r+s)} - \sqrt{(c+s)(r-s)}} < 1 \\
\Leftrightarrow \quad & \tfrac{\sqrt{c+s} - \sqrt{c-s}}{\sqrt{c+s} + \sqrt{c-s}} \tfrac{\sqrt{r+s} - \sqrt{r-s}}{\sqrt{r+s} + \sqrt{r-s}} < 1, \\
\Leftrightarrow \quad & \tfrac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \tfrac{\sqrt{\kappa_1} + 1}{\sqrt{\kappa_1} - 1} < 1
\end{aligned}
$$

and therefore $\eta_k \to 0$ as $k \to \infty$.

Appendix B

Detailed Pseudo Code for Algorithm PInvit

Algorithm 10 gives a pseudo code for PInvit is style of Algorithm 2 (p. 14). For the code presented here we assume that the action of P on a vector $\mathbf{x}$, so $P\mathbf{x}$, is not available. In the formulation of the algorithm we used the stopping tolerance $\tau = 0.1$. According to our experience $\tau = 0.8$ works fine, however with $\tau = 0.8$ the convergence area might be smaller than for $\tau = 0.1$. Further we point out that the pseudo code for PInvit, Algorithm 10, is a special case of Algorithm 4, generalized inexact inverse iteration, see page 55. As pointed out in Section 3.6 the important part is the linear solver, first benefiting from a 'better' right-hand side and second providing additonally to the soloution $\mathbf{y}^i$ the vector $P\mathbf{y}^i$.

Standard MINRES, see , for example the MatLab-routine, calculates a P-orthogonal basis for the Krylov subspace

$$\mathcal{K}_k(P^{-1}B, P^{-1}\mathbf{r}_0) \;=\; \text{span}(P^{-1}\mathbf{r}_0, (P^{-1}B)P^{-1}\mathbf{r}_0, \ldots, (P^{-1}B)^{k-1}P^{-1}\mathbf{r}_0),$$

where $P = P_1P_2$ denotes the preconditioner. Denoting the P-orthogonal basis by U we can write

$$P^{-1}AU \;=\; UT,$$

where T is the tridagonal matrix

$$U^TAU \;=\; (U^TP)P^{-1}AU \;=\; U^TPUT \;=\; T.$$

As this basis U is constructed iteratively we denote by $U_k = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$ its kth iterate. Now the solution $\mathbf{y}_k$ is given by $\mathbf{y}_k = \varphi U_k\mathbf{q}_k$ where $\mathbf{q}_k$ minimizes $\|\mathbf{e}_1 - T_k\mathbf{q}_k\|$ with $T_k = U_{k+1}^TAU_k$. Now let Q_kR_k be a QR-decomposition of T_k then we have $Q_k^T\mathbf{e}_1 - R_k\mathbf{q}_k$.

Algorithm 10: PInvit

- Given $\mathbf{x}^0$, σ^0,
- Solve with MINRES$^+$ $(A-\sigma^0 I)\mathbf{y}^0 = \mathbf{x}^0$ such that $\|\mathbf{x}^0 - (A-\sigma^0 I)\mathbf{y}^0\| \leq 0.1$,
- Update $\mathbf{x}^1 = \mathbf{y}^0 / \|\mathbf{y}^0\|$,
- For $i = 1, 2, 3, \ldots$
  - Calculate $\varrho^i = (\mathbf{x}^i)^T A \mathbf{x}^i$,
  - Set $\mathbf{b}^i := \mathbf{z}^{i-1} / \|\mathbf{z}^{i-1}\|$, where $\mathbf{z}^{i-1} = P\mathbf{y}^{i-1}$,
  - Solve with MINRES$^+$ $(A-\varrho^i I)\mathbf{y}^i = \mathbf{b}^i$ such that $\|\mathbf{b}^i - (A-\sigma^i I)\mathbf{y}^i\| \leq 0.1$,
  - Update $\mathbf{x}^{i+1} = \mathbf{y}^i / \|\mathbf{y}^i\|$,
  - Test for convergence

As T_k is tridiagonal, R_k is upper tridiagnonal and there exists a three term recurrence for R_k^{-1} which can be applied to any matrix. So, for example, $U_k R_k^{-1}$ uses the same recurrence formula on a different set of vectors.

In the pseudo-code for MINRES$^+$, see Algorithm 11, we use $\|P^{-1}\mathbf{w}\|_P$ but for any implementation we would use that $\|P^{-1}\mathbf{w}\|_P = \|\mathbf{w}\|_{P^{-1}} = (P^{-1}\mathbf{w}, \mathbf{w})$. The $\|\cdot\|_P$ is used here to indicate the P-orthonormality of U. The standard MINRES algorithm consists of the steps marked $\bullet$ and $-$ in Algorithm 11. So for MINRES$^+$ only the steps marked $+$ are added. Neither for standard MINRES nor for MINRES$^+$ do we need to store all basis vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots$, only the last three are needed. For the additional subspace we only need the new basis vector $\mathbf{u}_{k+1}^P$. The update sequence for $U_k R_k^{-1}$ uses three vectors, using the same sequence and three additional vectors, we obtain a three term update for $U_k^P R_k^{-1}$. With the addional vector for $\mathbf{z}_k = P\mathbf{y}_k$ MINRES$^+$ requires only five additional vectors compared with standard MINRES, and no further matrix vector products.

Algorithm 11: MINRES$^+$

- Given B, $\mathbf{b}$, τ and preconditioner P spd
- Set $\mathbf{u}_1 = P^{-1}\mathbf{b}/\ \|P^{-1}\mathbf{b}\|_P$,
+ Set $\mathbf{u}_1^P = \mathbf{b}/\ \|P^{-1}\mathbf{b}\|_P$,
- Calculate $\mathbf{w} = A\mathbf{u}_1 - \mathbf{u}_1(\mathbf{u}_1^T A\mathbf{u}_1)$,
- Set $\mathbf{u}_2 = P^{-1}\mathbf{w}/\ \|P^{-1}\mathbf{w}\|_P$,
+ Set $\mathbf{u}_2^P = \mathbf{w}/\ \|P^{-1}\mathbf{w}\|_P$,,
- Calculate $T_1 = U_2^T A U_1$ and $Q_1 R_1 = T_1$,
- Set $\mathbf{y}_1 = (U_1 R_1^{-1})Q_1^T \mathbf{e}_1$,
+ Set $\mathbf{z}_1 = (U_1^P R_1^{-1})Q_1^T \mathbf{e}_1$,
- Set $k = 1$
- Repeat until $\|\mathbf{b} - A\mathbf{y}_k\| \leq \tau$
  - Set $k = k + 1$,
  - Calculate $\mathbf{w}_k = A\mathbf{u}_k - \mathbf{u}_k(\mathbf{u}_k^T A\mathbf{u}_k) - \mathbf{u}_{k-1}(\mathbf{u}_{k-1}^T A\mathbf{u}_k)$,
  - Set $\mathbf{u}_{k+1} = P^{-1}\mathbf{w}_k/\ \|P^{-1}\mathbf{w}_k\|_P$,
  + Set $\mathbf{u}_{k+1}^P = \mathbf{w}_k/\ \|P^{-1}\mathbf{w}_k\|_P$,
  - Calculate $T_k = U_{k+1}^T A U_k$ and $Q_k R_k = T_k$,
  - Update $U_k R_k^{-1}$ and compute $\mathbf{y}_k = (U_k R_k^{-1})Q_k^T \mathbf{e}_1$,
  + Update $U_k^P R_k^{-1}$ and compute $\mathbf{z}_k = (U_k^P R_k^{-1})Q_k^T \mathbf{e}_1$.

Bibliography

Ablowitz, M. J. and A. S. Fokas (1997). *Complex Variables.* Cambridge University Press.

Absil, P.-A., R. Mahony, R. Sepulchre, and P. van Dooren (2002). A Grassmann-Rayleigh quotient iteration for computing invariant subspaces. *SIAM Review 44*(1), 57–73.

Berns-Müller, J., I. G. Graham, and A. Spence (2003). Inverse iteration and inexact solves. *Linear Algebra and its Apllications.* submitted.

Bouras, A. and V. Fraysee (2000). A relaxation strategy for the Arnoldi method in eigenproblems. Technical Report TR/PA/00/16, CERFACS.

Brown, P. N. and H. F. Walker (1997, January). GMRES on (nearly) singular systems. *SIAM Journal on Matix Analysis and Applications 18*(1), 37–51.

Campbell, S., I. Ipsen, C. Kelley, and C. Meyer (1996). GMRES and the minimal polynomial. *BIT 36*(4), 664–675.

Chapman, A. and Y. Saad (1997). Deflated and augmented Krylow subspace techniques. *Numerical Linear Algebra with Applications 4*(1), 43–66.

Chatelin, F. (1993). *Eigenvalues of Matrices.* John Wiley and Sons.

Embree, M. (1999). How descriptive are GMRES convergence bounds? Technical Report OUCL Numerical Analysis Group Technical Report 99/08, Oxford University Computing Laboratory.

Embree, M. (2003). The tortoise and the hare restart GMRES. *SIAM Review 45*(2).

Fischer, B. (1996). *Polynomial based iterative methods for symmetric linear systems.* Wiley.

Fischer, B. and R. W. Freund (1990). On the constrained Chebyshev approximation problem on ellipses. *Journal of Approximation Theory 62*, 297–315.

Fischer, B. and R. W. Freund (1991). Chebyshev polynomials are not always optimal. *Journal of Approximation Theory 65*, 261–272.

Fischer, B. and F. Peherstorfer (2001). Chebyshev approximation via polynomial mappings and the convergence behaviour of Krylov subspace methods. *ETNA 12*, 205–2-15.

Gantmacher, F. (1959a). *The Theory of Matrices*, Volume 1. Chelsea Publishing Company.

Gantmacher, F. (1959b). *The Theory of Matrices*, Volume 2. Chelsea Publishing Company.

Golub, G. H. and C. F. van Loan (1996). *Matrix Computations* (third ed.). John Hopkins.

Golub, G. H. and Q. Ye (2000). Inexact inverse iteration for generalized eigenvalue problems. *BIT 40*(4), 671–684.

Graham, I. G., A. Spence, and E. Vainikko (2003). Parallel iterative methods for Navier-Stokes equations and application to eigenvalue computation. *Concurrency and Computation: Practise and Experience.*

Greenbaum, A. (1997). *Iterative Methods for Solving Linear Systems.* SIAM.

Greenbaum, A., V. Pták, and Z. Strakoš (1996). Any noninctreasing convergence curve is possible for GMRES. *SIAM Journal on Matrix Analysis and Applications 17*(3), 465–469.

Hackbusch, W. (1994). *Iterative solution of large sparse systems of equations.* Springer-Verlag, Berlin.

Hawkins, S. (1999). *Computation of eigenvalues of large sparse matrices.* Ph. D. thesis, University of Bath,GB.

Ipsen, I. C. F. (1996). *Lineare Algebra and Analysis*, Volume 2, Chapter A history of inverse iteration, pp. 464–472. Walter de Gruyter Verlag Berlin.

Ipsen, I. C. F. (1998a). Expressions and bounds for the GMRES residual. *BIT 38*(2), 101–104.

Ipsen, I. C. F. (1998b). A note on the field of values of non-normal matrices. Technical Report CRSC-TR98-26. homepage:http://www4.nscu.edu/ ipsen/info.html.

Kelley, C. T. (1995). *Iterative Methods for Linear and Nonlinear Equations.* Frontiers in Applied Mathematics. SIAM.

Kilmer, M. and G. Stewart (1999). Iterative regularization and MINRES. *SIAM Journal on Matrix Analysis and Applications 21*(2), 613–628.

Knyazev, A. V. (2000). Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. Technical Report UCD-CCM 149, University of Colorado at Denver.

Knyazev, A. V. (2001). Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. Technical Report 2.

Knyazev, A. V. and K. Neymeyr (2003). A geometric theory for preconditioned inverse iteration. III: a short and sharp convergence estimate for generalized eigenvalue problems. *Linear Algebra and its Applications 358*, 95–114.

Lai, Y.-L., K.-Y. Lin, and L. Wen-Wei (1997). An inexact inverse iteration for large sparse eigenvalue problems. *Numerical Linear Algebra with Applications 1*, 1–13.

Liesen, J. (2000). Computable covergence bounds for GMRES. *SIAM Journal on Matrix Analysis and Applications 21*(3), 882–903.

Manteuffel, T. A. (1977). The Tchebychev iteration for nonsymmetric linear systems. *Numerische Mathematik 28*, 307–327.

Morgan, R. B. (1995, October). A restarted GMRES method augmented with eigenvectors. *SIAM Journal om Matrix Analysis and Applications 16*(4), 1254–1171.

Nachtigal, N. M., S. C. Reddy, and L. N. Trefethen (1992, July). How fast are nonsymmetric matrix iterations? *SIAM Journal on Matrix Analysis and Applications 13*(3), 778–795.

Neumaier, A. (1985). Residual inverse iteration for the nonlinear eigenvalue problem. *SIAM Journal on numerical Analysis 22*(5), 914–923.

Neymeyr, K. (2001a). A geometric theory for preconditioned inverse iteration i: Extrema of the Rayleigh quotient. *Linear Algebra and its Applications 322*, 61–85.

Neymeyr, K. (2001b). A geometric theory for preconditioned inverse iteration ii: Convergence estimates. *Linear Algebra and its Applications 322*, 87–104.

Neymeyr, K. (2002). A note on inverse iteration. preprint: Mathematishces Insititue der Universität Tübingen.

Notay, Y. (2003). Convergence analysis of inexact Rayleigh quotient iterations. *SIAM Journal on Matrix Analysis and Applications 24*, 627–644.

Ostrowski, A. (1957). On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. I. *Archive for Rational Mechanics and Analysis 1*, 233–241.

Ostrowski, A. (1958). On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. II. *Archieve for Rational Mechanics and Analysis 2*, 423–428.

Ostrowski, A. (1959). On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. V. *Archieve for Rational Mechanics and Analysis 3*, 472–481.

Ostrowski, A. (1960). On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. VI. *Archieve for Rational Mechanics and Analysis 4*, 153–165.

Paige, C. C. and M. Saunders (1975). Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis 12*, 617–629.

Parlett, B. N. (1980). *The Symmetric Eigenvalue Problem.* Prentice-Hall.

Rüde, U. and W. Schmid (1995, September). Inverse Multigrid Correction for Generalized Eigenvalue Computations. Technical report, Universität Augsburg.

Ruhe, A. and T. Wiberg (1972). The method of conjugate gradients used in inverse iteration. *BIT 12*, 543–554.

Saad, Y. (1993, March). A flexible inner-outer preconditioned GMRES algorithm. *SIAM Journal on Scientific Computing 14*(2), 461–469.

Saad, Y. (1996). *Iterative Methods for Sparse Linear Systems.* PWS Publishing Company.

Saad, Y. and M. H. Schultz (1986, July). GMRES: A generalized minimal residual algorithm for solving nonsymetric linear systems. *SIAM: Journal on Scientific and Statistical Computing 7*(3), 856–869.

Scott, D. S. (1981, February). Solving sparse symmetric generalized eigenvalue problems without factorization. *SIAM Journal on Numerical Analysis 18*(1), 102–110.

Simoncini, V. and L. Eldén (2002). Inexact Rayleigh quotient-type methods for eigenvalue computations. *BIT 42*(1), 159–182.

Simoncini, V. and D. B. Szyld (2002). Flexible inner-outer Krylov subspace methods. Technical Report IAN-CNR 1269, Istituto di Analisi Numerica, Pavia.

Simoncini, V. and D. B. Szyld (2003). Theorey of inexact Krylov subspace methods and applications to scientific computing. *SIAM journal on Scientific Computing*. to appear.

Sleijpen, G. L. G. and H. A. van der Vorst (2000). A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Review 42*(2), 267–293.

Smit, P. and M. H. C. Paardekooper (1999). The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra and its Applications 287*, 337–357.

Stewart, G. and J.-g. Sun (1990). *Matrix Perturbation Theory*. Academic Press.

Strang, G. (1986). *Introduction to Applied Mathematics*. Wellesley Cambridge Press.

Szyld, D. B. (1988). Criteria for combining inverse iteration and Rayleigh quotient iteration. *SIAM Journal on Numerical Analysis 25*(6), 1369–1375.

Trefethen, L. N. and D. Bau (1997). *Numerical Linear Algebra*. SIAM.

Turnbull, H. and A. Aitken (1932). *An Introduktion to the Theory of Canoical Matrices*. Blackie & Son Limited.

van der Vorst, H. and C. Vuik (1993). The superlinear convergence behavior of GMRES. *Journal of Computational and Applied Mathematics 48*, 327–341.

Wielandt, H. (1944). Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil V: Bestimmung höherer Eigenwerte durch gebrochene iteration. Technical Report B 44/J/37, Aerodynamische Versuchsanstalt Göttingen.

Wilkinson, J. (1958). The calculation of the eigenvectors of codiagonal matrices. *The Computer Journal 1*, 90–96.

Wilkinson, J. (1962). Calculation of the eigenvectors of a symmetric tridiagonal matrix by inverse iteration. *Numerische Mathematik 4*, 368–376.

Wilkinson, J. (1965). *The Algebraic Eigenvalue Problem*. Oxford Sciences Publications.

Zaslavsky, L. Y. (1995). An adaptive algebraic multigrid for reactor criticality calculations. *SIAM Journal on Scientific Computing 16*(4), 840–847.

STURMIAN NODAL SET ANALYSIS FOR HIGHER-ORDER PARABOLIC EQUATIONS AND APPLICATIONS

V.A. GALAKTIONOV

ABSTRACT. We describe local pointwise structure of multiple zeros of solutions of $2m$-th order uniformly parabolic equations ($m > 1$)

$$u_t = \sum_{|\beta| \le 2m} a_\beta(x,t) D_x^\beta u \quad \text{in } \mathbf{R}^N \times [-1,1],$$

with bounded continuous (for $|\beta| = 2m$) coefficients, in the existence-uniqueness class $\{|u(x,t)| \le B e^{b|x|^\alpha}\}$, where $B, b > 0$ are constants and $\alpha = 2m/(2m-1)$. Assuming that $u(0,0) = 0$, we perform a classification of all possible types of formation as $t \to 0^-$ and collapse as $t \to 0^+$ of multiple spatial zeros of the solutions $u(x,t)$. For one-dimensional second-order ($m = 1$) parabolic equations $u_t = a(x)u_{xx} + q(x)u$, this is known as Sturm's Theorem on zero sets established in 1836. In last twenty five years Sturm's ideas found new applications, generalizations and extensions in the parabolic PDE theory, mean curvature and curve shortening flows, symplectic geometry, etc.

Using such a local classification of multiple zeros, we establish a unique continuation theorem for higher-order parabolic PDEs and inequalities and estimate the Hausdorff dimension of nodal sets of solutions.

1. Introduction: main approach and results

Consider a general linear $2m$-th order parabolic equation with $m > 1$

$$u_t = \sum_{|\beta| \le 2m} a_\beta(x,t) D^\beta u \quad \text{in } Q_1 = \mathbf{R}^N \times [-1,1], \tag{1.1}$$

where the coefficients $\{a_\beta\}$ are real bounded for $|\beta| < 2m$ and real continuous for $|\beta| = 2m$ and satisfy the parabolicity condition: there exists a constant $\delta > 0$ such that

$$(-1)^m \sum_{|\beta|=2m} a_\beta(x,t)\xi^\beta \le -\delta|\xi|^{2m} \quad \text{for all } (x,t) \in Q_1 \text{ and } \xi \in \mathbf{R}^N. \tag{1.2}$$

Let $u(x,t)$ be a classical, $C^{2m,1}$, solution of (1.1) in the existence-uniqueness class of locally measurable functions

$$\mathcal{U} = \{|u(x,t)| \le C e^{c|x|^\alpha}\}, \quad \text{with the exponent } \alpha = 2m/(2m-1), \tag{1.3}$$

where C, c are positive constants; see the classical parabolic theory, [6], [7], [16].

Date: August 29, 2003.

1991 *Mathematics Subject Classification.* 35K55, 35K65.

Key words and phrases. Higher-order parabolic equations, multiple zeros, asymptotic behaviour, non self-adjoint operator, spectrum, nodal sets.

Research supported by RTN network HPRN-CT-2002-00274 and CERN-INTAS00-0136.

1.1. **Sturmian backward rescaling and "micro-structure" of the PDE.** The main goal of the paper is to detect the local pointwise structure of the solutions at any fixed internal point in Q_1 using the *optimal parabolic blow-up* $\{x,t\}$-scaling. In general, this is quite a delicate *asymptotic* problem, and even among canonical linear PDEs, there are just a few examples related to the *heat equation* with known pointwise structure of solutions (references are given below). Indeed, such an analysis can play a special role in the theory of linear and nonlinear PDEs. Indeed, once such a local pointwise structure of solutions is known, including description of all possible "singularities", the existence-uniqueness theory can be produced by fixing those functional settings which exclude those structures and singularities that can violate the desired "regularity" and uniqueness of solutions. Such a pointwise approach to PDEs differs from another direction of the PDE theory devoted to statistical, stochastic and chaotic ("turbulent") properties of general solution subsets. Here other probability and averaging methods apply having, nevertheless, some not that straightforward connections with the micro-scale, pointwise structure of solutions.

In mechanics and physics, dealing with continuous media like fluids, gases or porous media, the questions of the pointwise behaviour are often called as the problems of "micro-structure" (or the "turbulent", molecular one in fluid dynamics) of the medium which eventually determine the global coherent patterns occurring via the given turbulent mechanism. In the present case, the "medium" is prescribed by solutions of the PDE under consideration, and therefore, loosely speaking, we are going to study the internal "molecular" structure of the given class of PDEs.

To this end, we consider the finite-time asymptotic behaviour of the solution at the origin $(0,0)$ using the *Sturmian backward variable* (see related references below)

$$y = x/(-t)^{1/2m}, \quad (x,t) \in Q_1^- = \mathbf{R}^N \times [-1,0), \tag{1.4}$$

with the blow-up at $t = 0^-$. We next introduce the corresponding new time variable

$$\tau = -\ln(-t) \to \infty \quad \text{as} \ \ t \to 0^-. \tag{1.5}$$

Note that as $t \to 0^-$, i.e., as $\tau \to \infty$, the behaviour of solutions on compact subsets in y, $|y| \le C = \text{const}$, implies a natural parabolic "zoom" on fast shrinking subsets in the original spatial variable x, $|x| \le C(-t)^{1/2m} \equiv Ce^{-\tau/2m} \to 0$. The rescaled variable (1.4) is purely dimensional for any $2m$-th order linear or quasilinear uniformly parabolic equation, and, as we will show, no non-trivial solution structures take place on smaller compact subsets in $\{y,\tau\}$. Actually, these rescaled variables define the optimal scales of the non-trivial "turbulent" behaviour available in the PDEs (1.1) with sufficiently smooth coefficients.

In order to describe the micro-structure provided by the PDE, we study formation and collapse of *multiple zeros* of a given classical solution $u(x,t)$ of (1.1) assuming that

$$u(0,0) = 0. \tag{1.6}$$

We then need to perform suitable asymptotic analysis of the behaviour of the solution as $t \to 0^-$ and, in the next step, as $t \to 0^+$. The first limit is of crucial importance and actually determines all possible types of multiple zeros which can occur in the parabolic

equation (1.1). This gives the variety of such micro-patterns and hence explains the degree of local "turbulence" available.

Since $D_x^\beta u \equiv (-t)^{-|\beta|/2m} D_y^\beta u$, in terms of the new independent variables $\{y, \tau\}$, the solution $u = u(y, \tau)$ satisfies the *rescaled equation*

$$u_\tau = \mathbf{B}^* u + \mathbf{C}(\tau) u \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+, \tag{1.7}$$

where $\mathbf{B}^*$ is the $2m$-th order elliptic operator

$$\mathbf{B}^* = \textstyle\sum_{|\beta|=2m} A_\beta D_y^\beta - \frac{1}{2m} y \cdot \nabla, \quad \text{and } A_\beta = a_\beta(0,0). \tag{1.8}$$

For convenience, $\mathbf{B}^*$ is written as adjoint to another operator $\mathbf{B}$ given below. The higher-order principle counterpart

$$\mathbf{B}_0 = \textstyle\sum_{|\beta|=2m} A_\beta D_y^\beta$$

is a symmetric homogeneous $2m$-th order elliptic operator with constant coefficients.

The time-dependent perturbation $\mathbf{C}(\tau)$ in (1.7) is given by

$$\mathbf{C}(\tau) = \textstyle\sum_{|\beta|=2m} R_\beta(y,\tau) D_y^\beta + \sum_{|\beta|<2m} e^{-(2m-|\beta|)\tau/2m} a_\beta(y e^{-\tau/2m}, -e^{-\tau}) D_y^\beta, \tag{1.9}$$

$$R_\beta(y,\tau) \equiv a_\beta(y e^{-\tau/2m}, -e^{-\tau}) - a_\beta(0,0).$$

Therefore, $\mathbf{C}(\tau)$ is *exponentially small* if coefficients $\{a_\beta\}$ are continuous for all $|\beta| = 2m$ and are uniformly bounded for any $|\beta| < 2m$. These are main necessary assumptions on the PDE coefficients. Then, as $\tau \to \infty$, uniformly on compact subsets,

$$R_\beta(y,\tau) = O(e^{-\tau/2}). \tag{1.10}$$

It follows that on smooth solutions, $\mathbf{C}(\tau)u$ in (1.7) is an exponentially small perturbation satisfying as $\tau \to \infty$, uniformly on compact subsets,

$$|\mathbf{C}(\tau)u| = O(e^{-\tau/2m}). \tag{1.11}$$

Further estimates of perturbations are to be performed in the weighted Sobolev spaces associated with operator (1.8) and the adjoint one to be introduced next.

1.2. **Linear non self-adjoint operators.** It follows from equation (1.7) that, first, one needs to study spectral and other properties of the linear operator $\mathbf{B}^*$ (and of the adjoint one $\mathbf{B}$). Section 2 is devoted to some preliminaries concerning the fundamental solutions of operators $\partial/\partial - \mathbf{B}$ and $\partial/\partial - \mathbf{B}^*$, semigroups $e^{\mathbf{B}\tau}$, $e^{\mathbf{B}^*\tau}$ and resolvents of $\mathbf{B}$ and $\mathbf{B}^*$. Here we study the unperturbed homogeneous parabolic equation

$$u_\tau = \mathbf{B}^* u \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+. \tag{1.12}$$

In Section 3 we describe the spectral properties the adjoint non self-adjoint operator

$$\mathbf{B} = \textstyle\sum_{|\beta|=2m} A_\beta D_y^\beta + \frac{1}{2m} y \cdot \nabla + \frac{N}{2m} I$$

in the weighted space $L^2_\rho(\mathbf{R}^N)$, where

$$\rho(y) = e^{a|y|^\alpha} > 0 \quad \text{and } a > 0 \text{ is a sufficiently small constant.} \tag{1.13}$$

We show that $\mathbf{B}$ has the point spectrum only $\sigma(\mathbf{B}) = \{\lambda_\beta = -|\beta|/2m\}$, and the eigenfunctions $\Phi = \{\psi_\beta\}$ form a complete subset in L^2_ρ. In Section 4 we study spectral properties

of the adjoint operator $\mathbf{B}^*$ in $L^2_{\rho^*}(\mathbf{R}^N)$ with $\rho^* = 1/\rho$ and describe the complete subset $\Phi^* = \{\psi^*_\beta\}$ of polynomial eigenfunctions. In Section 5 we describe subspaces where eigenfunction subsets Φ and Φ^* are closed.

1.3. **Main results: formation of multiple zeros.** Using eigenfunction expansions, we show that multiple zeros at the origin $(0,0)$ of any suitable solution $u(x,t) \not\equiv 0$ of (1.7) has a local structure corresponding to stable subspace of $\mathbf{B}^*$. Namely, in Section 6, we show that for any such solution, there exists a finite $l > 0$ such that as $\tau \to \infty$,

$$u(y,\tau) = e^{-l\tau/2m}[\varphi^*_l(y) + o(1)] \quad \text{uniformly on compact subsets,} \tag{1.14}$$

where φ^*_l is a polynomial eigenfunction of $\mathbf{B}^*$,

$$\varphi^*_l(y) = \textstyle\sum_{|\beta|=l} c_\beta \psi^*_\beta(y) \not\equiv 0, \tag{1.15}$$

corresponding to the eigenvalue $\lambda_l = -l/2m < 0$. Therefore, (1.14) describes all possible types of formation of multiple zeros for uniformly parabolic PDEs (1.1), so that any blow-up formation of an l-th order multiple zero at $(0,0)$ after rescaling (1.4) is driven as $t \to 0^-$ by zero surfaces of an l-th order polynomial eigenfunction (1.15) of $\mathbf{B}^*$.

1.4. **On second-order parabolic equations.** The zero formation analysis is classical for $m = 1$. In one dimension it was performed by C. Sturm in 1836 [18] for C^∞ solutions of linear parabolic equations $u_t = a(x)u_{xx} + q(x)u$. For the *heat equation* in $\mathbf{R}^N$ $(m = 1)$

$$u_t = \Delta u \quad \text{in } Q_1^-, \tag{1.16}$$

introducing Sturm's variable (1.4), $y = x/(-t)^{1/2}$, yields the rescaled equation (1.12) with the second-order symmetric operator

$$\mathbf{B}^* = \Delta - \tfrac{1}{2} y \cdot \nabla \equiv \tfrac{1}{\rho^*} \nabla \cdot (\rho^* \nabla), \quad \text{where } \rho^*(y) = e^{-|y|^2/4}. \tag{1.17}$$

In this case $\mathbf{B}^*$ is known to be self-adjoint in $L^2_{\rho^*}$ with the domain $H^2_{\rho^*}$. It has the discrete spectrum $\sigma(\mathbf{B}^*) = \{\lambda_\beta = -|\beta|/2\}$ and the resolvent is a compact integral operator. The eigenfunctions $\{\psi_\beta\}$, which are Hermite polynomials $c_\beta H_\beta$ in $\mathbf{R}^N$ [2] and c_β are normalization multipliers, given by the generating formula $D^\beta e^{-|y|^2/4} = H_\beta(y) e^{-|y|^2/4}$, form an orthonormal basis in $L^2_{\rho^*}$. Since for $m = 1$, (1.17) is self-adjoint, the classical Agmon-Ogawa estimates apply to the corresponding perturbed rescaled equations like (1.7) to ensure the convergence (1.14). We refer to a detailed analysis for $m = 1$ in [3].

1.5. **Applications: unique backward continuation and some global properties of nodal sets.** Using the optimal characterization (1.14) of arbitrary multiple zeros for solutions $u(x,t) \not\equiv 0$, in Section 7 we establish a unique backward continuation theorem for $2m$-th order parabolic equations.

In Section 8 we study some global properties of the nodal set

$$Z_t[u] = \{x \in \mathbf{R}^N : \ u(x,t) = 0\}, \ \ t \in (-1,1), \tag{1.18}$$

of nontrivial solutions to (1.1). We prove that the Hausdorff dimension of $Z_t[u]$ satisfies

$$\dim_{\mathcal{H}} Z_t[u] \le N - 1. \tag{1.19}$$

2. Preliminaries: fundamental solution, semigroups, resolvents

2.1. **Fundamental solution.** Consider the Cauchy problem for the homogeneous $2m$-th order parabolic equation with constant coefficients

$$u_t = \mathbf{B}_0 u \equiv \textstyle\sum_{|\beta|=2m} A_\beta D_x^\beta u \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+, \quad u(x,0) = u_0(x) \in \mathcal{U} \cap \{t=0\}. \tag{2.1}$$

Let $b(x,y)$ be the fundamental solution of the operator $\partial/\partial t - \mathbf{B}_0$, [6, 7]. It has the self-similar form

$$b(x,t) = t^{-N/2m} f(y), \quad y = x/t^{1/2m}, \tag{2.2}$$

where f is a unique solution of the linear elliptic equation

$$\mathbf{B}f \equiv \mathbf{B}_0 f + \tfrac{1}{2m} y \cdot \nabla f + \tfrac{N}{2m} f = 0 \quad \text{in } \mathbf{R}^N, \quad \int_{\mathbf{R}^N} f = 1. \tag{2.3}$$

The following estimates holds [6]:

$$|f(y)| \le D e^{-d|y|^\alpha} \quad \text{in } \mathbf{R}^N, \tag{2.4}$$

where D, d are positive constants. The unique solution of (2.1) is given by the convolution

$$u(x,t) = b(t) * u_0 \equiv t^{-N/2m} \int_{\mathbf{R}^N} f((x-z)t^{-1/2m}) u_0(z) dz. \tag{2.5}$$

2.2. **Semigroup with infinitezimal generator B.** The rescaled solution

$$w(y,\tau) = t^{N/2m} u(yt^{1/2m}, t), \quad \text{where } \tau = \ln t \in \mathbf{R}, \tag{2.6}$$

satisfies the parabolic equation

$$w_\tau = \mathbf{B} w. \tag{2.7}$$

One can see that $w(y,\tau)$ satisfies the Cauchy problem for (2.7) in $\mathbf{R}^N \times \mathbf{R}_+$ with initial data at $\tau = 0$ (i.e., at $t = 1$)

$$w_0(y) = u(y,1) \equiv b(1) * u_0. \tag{2.8}$$

Rescaling convolution (2.5) yields the following explicit representation of the semigroup with the infinitezimal generator $\mathbf{B}$:

$$w(y,\tau) = e^{\mathbf{B}\tau} w_0 \equiv \int_{\mathbf{R}^N} f(y - \zeta e^{-\tau/2m}) u_0(\zeta) d\zeta, \quad \tau \ge 0. \tag{2.9}$$

Performing another rescaling $w(y,\tau) = (1+t)^{N/2m} u(y(1+t)^{1/2m}, t)$ with the new time variable $\tau = \ln(1+t) : \mathbf{R}_+ \to \mathbf{R}_+$, we obtain the solution $w(y,\tau)$ of the Cauchy problem for equation (2.7) with initial data $w_0(y) \equiv u_0(y)$. Rescaling (2.5), we deduce a standard (without the relation (2.8)) representation of the semigroup

$$w(y,\tau) = e^{\mathbf{B}\tau} w_0 \equiv (1-e^{-\tau})^{-N/2m} \int f((y - \zeta e^{-\tau/2m})(1-e^{-\tau})^{-1/2m}) w_0(\zeta) d\zeta. \tag{2.10}$$

2.3. **Semigroup with the adjoint infinitezimal generator $\mathbf{B}^*$.** In order to construct the semigroup with the infinitezimal generator $\mathbf{B}^*$, we introduce the rescaled variables corresponding to blow-up as $t \to 1^-$,

$$u(x,t) = w(y,\tau), \quad y = x/(1-t)^{1/2m}, \quad \tau = -\ln(1-t) : (0,1) \to \mathbf{R}_+.$$

Then w solves the problem

$$w_\tau = \mathbf{B}^* w \quad \text{for } \tau > 0, \quad w(0) = u_0. \tag{2.11}$$

Rescaling solution (2.5), we obtain the following representation of the semigroup:

$$w(y,\tau) = e^{\mathbf{B}^*\tau} w_0 \equiv (1-e^{-\tau})^{-N/2m} \int f((ye^{-\tau/2m} - \zeta)(1-e^{-\tau})^{-1/2m}) u_0(\zeta) d\zeta. \tag{2.12}$$

2.4. **Resolvents.** Using the descent method for constructing of resolvents, [6], fixing $\lambda \in \mathbb{C}$, we consider an auxiliary non-homogeneous problem $w_\tau = \mathbf{B}w - e^{\lambda\tau} g$ for $\tau > 0$ with $w(0) = 0$. Here we assume that g belongs to the weighted L^2-space $L^2_\rho(\mathbf{R}^N)$, see the next section. Performing formal computations and setting $w = e^{\lambda\tau} v$ yields the equation $v_\tau = (\mathbf{B} - \lambda I)v - g|$ and hence $v(\tau) = -\int_0^\tau e^{(\mathbf{B}-\lambda I)(\tau-s)} g ds$. Setting $\tau - s = \eta$ and passing to the limit $\tau \to \infty$ yield that there exists a limit

$$v(\infty) = -\int_0^\infty e^{(\mathbf{B}-\lambda I)\eta} g \, d\eta \equiv (\mathbf{B} - \lambda I)^{-1} g$$

provided that the integral converges. Using the semigroup representation (2.10) and changing the variable $e^{-\eta} = z \in (0,1)$ yield the integral operator

$$(\mathbf{B} - \lambda I)^{-1} g = \int_{\mathbf{R}^N} K(y,\zeta) g(\zeta) d\zeta \quad \text{with the kernel} \tag{2.13}$$

$$K(y,\zeta) = -\int_0^1 z^{\lambda-1} (1-z)^{-N/2m} f((y - \zeta z^{1/2m})(1-z)^{-1/2m}) dz. \tag{2.14}$$

Similarly, representation of the resolvent of the adjoint operator $\mathbf{B}^*$ is

$$(\mathbf{B}^* - \lambda I)^{-1} g = \int_{\mathbf{R}^N} K^*(y,\zeta) g(\zeta) d\zeta, \quad \text{where} \tag{2.15}$$

$$K^*(y,\zeta) = -\int_0^1 z^{\lambda-1} (1-z)^{-N/2m} f((yz^{1/2m} - \zeta)(1-z)^{-1/2m}) dz. \tag{2.16}$$

Both operators (2.13) and (2.15) are compact for $\lambda \notin \sigma(\mathbf{B})$, see below.

3. Spectral properties of B

It is convenient to begin with spectral properties of operator (2.3) which for $m > 1$ is not symmetric and does not admit a self-adjoint extension in any weighted space $L^2_\rho = L^2_\rho(\mathbf{R}^N)$. As a differential operator with smooth coefficients, it is closable [8]. We consider $\mathbf{B}$ in the weighted space L^2_ρ with the exponentially growing weight function (1.13), where $a > 0$ is a small positive constant and, at least,

$$a < 2d. \tag{3.1}$$

The scalar product in L^2_ρ is denoted by $\langle\cdot,\cdot\rangle_\rho$ and $\|\cdot\|_\rho$ is the induced norm. Note that $m=1$ is the only case where the operator adjoint to (1.17) is symmetric in a weighted L^2-space and admits a unique Friedrichs self-adjoint extension [2]. Indeed,

$$\mathbf{B}=\Delta+\tfrac{1}{2}y\cdot\nabla+\tfrac{N}{2}I\equiv\tfrac{1}{\rho}\nabla\cdot(\rho\nabla)+\tfrac{N}{2}I, \tag{3.2}$$

where $\rho(y)=e^{|y|^2/4}$ is the inverse Gaussian kernel, i.e., (1.13) with $\alpha=2$ and $a=1/4$. Note that if, according to (3.1), $a<1/2$ but $a\neq 1/4$, then (3.2) is not self-adjoint in L^2_ρ but nevertheless enjoys a number of good properties which are shown to remain valid for any $m>1$.

The crucial spectral and various other properties of eigenvalues and eigenfunctions of $\mathbf{B}$ can be obtained directly from the explicit representation of the (analytic) semigroup (2.9) or (2.10), which, indeed, is a great advantage of the analysis.

3.1. **Domain of the operator.** Consider a Hilbert space of functions H^{2m}_ρ with the inner product and the norm

$$\langle v,w\rangle_{2m,\rho}=\int_{\mathbf{R}^N}\rho\textstyle\sum_{k=0}^{2m}D^kv\,D^kw\,dy,\quad \|v\|^2_{2m,\rho}=\int_{\mathbf{R}^N}\rho\sum_{k=0}^{2m}|D^kv|^2dy, \tag{3.3}$$

where D^kv denote vectors $\{D^\beta v,\ |\beta|=k\}$. Obviously, $H^{2m}_\rho\subset L^2_\rho\subset L^2$.

Proposition 3.1. $\mathbf{B}: H^{2m}_\rho\to L^2_\rho$ *is a bounded linear operator.*

Proof. It follows from (2.3) that $\mathbf{B}v\in L^2_\rho$ for any $v\in H^{2m}_\rho$ provided that

$$\int_{\mathbf{R}^N}\rho|y\cdot\nabla v|^2dy\le C\|v\|^2_{2m,\rho}\quad\text{for any } v\in H^{2m}_\rho,\quad C=\text{const}>0. \tag{3.4}$$

The proof follows the lines of a similar analysis in [5], Section 2. □

Embeddings like (3.4) are associated with the well known general estimates in weighted spaces (see p. 40 in Maz'ja's book [15] and Lemma 2.1 in [10]), which go back to the classical Hardy inequality established in 1920, [9].

3.2. **Discrete spectrum.**

Lemma 3.1. *(i) The spectrum of* $\mathbf{B}$ *consists of real eigenvalues only,*

$$\sigma(\mathbf{B})=\{\lambda_\beta=-|\beta|/2m,\ |\beta|=0,1,2,...\}, \tag{3.5}$$

and eigenvalues λ_β *have finite multiplicity with eigenfunctions*

$$\psi_\beta(y)=(-1)^{|\beta|}(\beta!)^{-1/2}D^\beta f(y). \tag{3.6}$$

(ii) The eigenfunction subset $\Phi=\{\psi_\beta\}$ *is complete in* L^2 *and in* L^2_ρ.

(iii) Resolvent $(\mathbf{B}-\lambda I)^{-1}$ *is compact in* L^2_ρ *for any* $\lambda\notin\sigma(\mathbf{B})$.

In the case $m=1$, for operator (3.2) we have that f is the positive rescaled Gaussian kernel $f(y)=(4\pi)^{-N/2}e^{-|y|^2/4}$, and the eigenfunctions are

$$\psi_\beta(y)=c_\beta e^{-|y|^2/4}H_\beta(y),\quad H_\beta(y)\equiv H_{\beta_1}(y_1)...H_{\beta_N}(y_N),$$

where H_β denote separable Hermite polynomials in $\mathbf{R}^N$. Operator $\mathbf{B}$ with the domain H^2_ρ, where $\rho = e^{|y|^2/4}$, is self-adjoint and the eigenfunctions form an orthonormal basis in L^2_ρ [2], p. 48. For $m > 1$, the eigenfunctions are orthogonal to the adjoint ones in terms of the dual, L^2-product. The adjoint eigenfunctions are polynomials which form a complete subset in $L^2_{\rho^*}$ with decaying exponential weight $\rho^*(y) = 1/\rho(y) = e^{-a|y|^\alpha}$; see the next section.

Proof. (i) Spectrum and eigenfunctions. Let $l = |\beta|$. The existence of such eigenvalues and eigenfunctions follows by applying D^β to the elliptic equation (2.3)

$$D^\beta \mathbf{B} f \equiv \mathbf{B} D^\beta f + \tfrac{|\beta|}{2m} D^\beta f = 0. \tag{3.7}$$

In order to show that $\mathbf{B}$ admits no other eigenvalues, we consider the explicit semigroup representation (2.9). Using Taylor power series of the analytic kernel (convergence of such series is studied in Section 4)

$$f(y - ze^{-\tau/2m}) = \textstyle\sum_{(\beta)} e^{-|\beta|\tau/2m} \dfrac{(-1)^{|\beta|}}{\beta!} D^\beta f(y) z^\beta \equiv \sum_{(\beta)} e^{-|\beta|\tau/2m} \dfrac{1}{\sqrt{\beta!}} \psi_\beta(y) z^\beta, \tag{3.8}$$

where $z^\beta \equiv z_1^{\beta_1} ... z_N^{\beta_N}$, and substituting it into (2.9), we arrive at the following eigenfunction expansion of the solution:

$$w(y,\tau) = \textstyle\sum_{(\beta)} e^{-|\beta|\tau/2m} M_\beta(u_0) \psi_\beta(y), \tag{3.9}$$

where $\lambda_\beta = -|\beta|/2m$ and $\psi_\beta(y)$ are the eigenvalues and eigenfunctions of $\mathbf{B}$. Here

$$M_\beta(u_0) = (\beta!)^{-1/2} \int_{\mathbf{R}^N} z^\beta u_0(z) dz \tag{3.10}$$

are the momenta of the initial datum w_0 (recall the relation (2.8) between w_0 and u_0).

Let $\langle \cdot, \cdot \rangle$ be the dual inner product in L^2. Then $M_\beta(u_0) = (\beta!)^{-1/2} \langle z^\beta, u_0 \rangle \equiv \langle w_0, \psi^*_\beta \rangle$, where ψ^*_β are polynomial eigenfunctions of the adjoint operator $\mathbf{B}^*$ to be described in the next section. It follows from the asymptotic analysis of expansion (3.9) as $\tau \to \infty$ that no other eigenfunctions exist, all eigenvalues are real and are given in (3.5).

(ii) Completeness. Firstly, let us show that the system of the eigenfunctions $\{D^\beta f\}$ is complete in L^2. By the Riesz-Fischer theorem, we have to show that, given a function $g \in L^2$, the equalities

$$\int D^\beta f(x) g(x) dx = 0 \quad \text{for any } \beta \tag{3.11}$$

imply that $g = 0$. Let $\hat f(\xi)$ and $\hat g(\xi)$ be the Fourier transforms of f and g. Then

$$\int \xi^\beta \hat f(\xi) \hat g(-\xi) d\xi = 0 \quad \text{for any } \beta.$$

Applying the Fourier transform to equation (2.3) yields $P_{2m}(\xi)\hat f - \frac{1}{2m} \xi \cdot \nabla \hat f = 0$, where by the parabolicity condition (1.2)

$$P_{2m}(\xi) = \textstyle\sum_{|\beta|=2m} A_\beta (i\xi)^\beta = (-1)^m \sum_{|\beta|=2m} A_\beta \xi^\beta \le -\delta |\xi|^{2m} \quad \text{in } \mathbf{R}^N. \tag{3.12}$$

Since $P_{2m}(\xi)$ is a homogeneous $2m$-th order polynomial, by Euler's formula $\xi \cdot \nabla P_{2m}(\xi) = 2mP_{2m}(\xi)$ we find that (recall that $b(x,0) = \delta(x)$)

$$\hat{f}(\xi) \equiv \mathcal{F}(f(\cdot))(\xi) = e^{P_{2m}(\xi)} \implies \int \xi^\beta e^{P_{2m}(\xi)} \hat{g}(-\xi) d\xi = 0 \quad \text{for any } \beta. \tag{3.13}$$

By the parabolicity condition (1.2), the function $M(z) = \int e^{P_{2m}(\xi)} \hat{g}(-\xi) e^{iz\xi} d\xi$ is entire analytic in $\mathbb{C}^N$ (since $|e^{iz\xi}| \le e^{|\mathrm{Im}\, z||\xi|}$). Equality (3.13) means that $D^\beta M(0) = 0$ for any β. Therefore, $M(z) \equiv 0$. Thus, $\hat{g}(\xi) = 0$ almost everywhere and $g = 0$.

Secondly, in order to prove completeness in L^2_ρ, as in [5], we suppose that a function $g \in L^2_\rho$ is orthogonal relative to the inner product in L^2_ρ to all eigenfunctions, i.e.,

$$\int \rho(y) D^\alpha f(y) g(y) dy = 0 \quad \text{for all } \alpha.$$

Since f is analytic, it implies that $\int \rho(y) f(y-x) g(y) dy = 0$ for all $x \in \mathbf{R}^N$. Consider the Cauchy problem for the linear parabolic equation (2.1) with initial data $u_0(x) = \rho(x) g(x)$. One can see from the Poisson-type integral (2.5) and (2.4) by using Eidel'man's estimate [6], Lemma 5.1 (see also an extension for integrals over $\mathbf{R}^N$ in [5], Proposition 4.1) that the solution exists for all $t \ge 1$, provided that the exponent $a > 0$ in the weight (1.13) satisfies (3.1). Then $u(x,t)$ is analytic in x. We have $u(x,1) = \int f(x-y) g(y) \rho(y) dy$. Therefore, $u(x,1) \equiv 0$. It follows by the uniqueness theorem for the inverse parabolic equation [7], p. 181, that $u(x,0) = 0$, and $g = 0$.

(iii) *Compact resolvent.* We next deduce that $(\mathbf{B} - I)^{-1}$ is an integral compact operator and has a point spectrum only. The proof is similar to that in [5], Theorem 2.2. An simpler compactness analysis in a subspace of L^2_ρ is presented in Section 5.

This completes the proof of Lemma 3.1. □

4. Discrete spectrum and polynomial eigenfunctions of $\mathbf{B}^*$

Let us describe the eigenfunctions of the adjoint operator (1.8). We consider $\mathbf{B}^*$ in the weighted space $L^2_{\rho^*}$ with the exponentially decaying weight function

$$\rho^*(y) \equiv 1/\rho(y) = e^{-a|y|^\alpha} > 0, \tag{4.1}$$

and ascribe to $\mathbf{B}^*$ the domain $H^{2m}_{\rho^*}$ dense in $L^2_{\rho^*}$. Then $\mathbf{B}^* : H^{2m}_{\rho^*} \to L^2_{\rho^*}$ is adjoint to $\mathbf{B}$,

$$\langle \mathbf{B}v, w \rangle = \langle v, \mathbf{B}^* w \rangle \quad \text{for any } v \in H^{2m}_\rho, \ w \in H^{2m}_{\rho^*}, \tag{4.2}$$

and hence is a bounded linear operator, [13], Chapt. 4.

4.1. Discrete spectrum.

Let us fix the main spectral properties of $\mathbf{B}^*$.

Lemma 4.1. *(i) The spectrum of* $\mathbf{B}^*$ *is discrete,*

$$\sigma(\mathbf{B}^*) = \sigma(\mathbf{B}) = \{\lambda_\beta = -|\beta|/2m, \ |\beta| = 0, 1, 2, ...\}, \tag{4.3}$$

and eigenfunctions $\{\psi^*_\beta(y)\}$ *are polynomials of order* $|\beta|$,

$$\psi^*_\beta(y) = (\beta!)^{-1/2} \left[y^\beta + \textstyle\sum_{j=1}^{[|\beta|/2m]} \frac{1}{j!} (-\mathbf{B}_0)^j y^\beta \right]. \tag{4.4}$$

(ii) The eigenfunction subset $\Phi^* = \{\psi^*_\beta\}$ *is complete in* $L^2_{\rho^*}$.
(iii) Resolvent $(\mathbf{B}^* - \lambda I)^{-1}$ *is compact in* $L^2_{\rho^*}$ *for any* $\lambda \notin \sigma(\mathbf{B}^*)$.

Proof. (i) *Spectrum and polynomial eigenfunctions.* Firstly, $\sigma(\mathbf{B}) = \sigma(\mathbf{B}^*)$ [13]. Secondly, let us prove that $\{\psi^*_\beta\}$ are polynomials. Let $V(\xi) = \int \psi^*(y) e^{-iy\cdot\xi} dy$ be the Fourier transform of an eigenfunction, $\mathbf{B}^*\psi^* = \lambda\psi^*$. Then V solves the first-order equation

$$\tfrac{1}{2m}\xi \cdot \nabla V + (\tfrac{N}{2m} + P_{2m}(\xi))V = \lambda V \quad \text{in } \mathbf{R}^N. \tag{4.5}$$

The general solution is given by

$$V(\xi) = \Phi|\xi|^{2m\lambda - N} e^{-P_{2m}(\xi)} \quad \text{in} \quad \mathbf{R}^N \setminus \{0\}, \tag{4.6}$$

where $\Phi = \Phi(\xi/|\xi|)$ is an arbitrary smooth function on the unit sphere $S_1 = \{|\xi| = 1\}$ in $\mathbf{R}^N$. In view of the parabolicity assumption (1.2), we obtain in (4.6) an exponentially growing factor $|e^{-P_{2m}(\xi)}| \geq e^{\delta|\xi|^{2m}}$ as $\xi \to \infty$. Therefore, the only distributions satisfying equation (4.5) correspond to $\Phi \equiv 0$ on S_1, i.e., those having supports concentrated at the origin $\xi = 0$. Therefore, $\psi^*_\beta(y)$ must be a polynomial. If its degree is k, then

$$\psi^*(y) = \textstyle\sum_{j=0}^{s} P_j(y) \quad \text{with } \ s = [k/2m], \tag{4.7}$$

where $P_j(y)$ are homogeneous polynomials of degree $k - 2mj$. Since by the Euler identity $-\frac{1}{2m}\sum_{j=1}^{N} y_j \partial P_0(y)/\partial y_j = -\frac{k}{2m} P_0(y) = \lambda P_0(y)$, we see that $\lambda = -k/2m$ and, hence, $P_0(y)$ is an arbitrary homogeneous polynomial of degree k. Other polynomials $P_j(y)$ are then defined as follows:

$$P_j(y) = (j!)^{-1}(-\mathbf{B}_0)^j P_0(y), \quad j = 1, ..., s. \tag{4.8}$$

We fix $P_0(y) = y^\beta/\sqrt{\beta!}$ in (4.7), so that for eigenfunctions (3.6) of $\mathbf{B}$, the corresponding adjoint eigenfunctions take the form (4.4). Then the orthonormality condition holds

$$\langle \psi_\beta, \psi^*_\gamma \rangle = \delta_{\beta,\gamma} \quad \text{for any } \ \beta \text{ and } \gamma, \tag{4.9}$$

where $\delta_{\beta,\gamma}$ is the Kronecker delta. Note that operators $\mathbf{B}$ and $\mathbf{B}^*$ have zero Morse index and do not have eigenvalues with positive real parts. For $\beta = 0$ the eigenfunctions are

$$\psi_0(y) = f(y), \quad \psi^*_0(y) = 1, \tag{4.10}$$

so that $\langle \psi_0, \psi^*_0 \rangle = 1$ by the definition (2.3) of the fundamental solution.

(ii) *Completeness.* It follows from the well-known fact that polynomials $\{y^\beta\}$, which are higher-order terms in any eigenfunction ψ_β, are complete in suitable weighted L^p-spaces; see [13], p. 431. Then (4.7) implies the completeness of Φ^* in $L^2_{\rho^*}$.
(iii) *Compact resolvent.* Since $(\cdot)^*$ and $(\cdot)^{-1}$ commute for operators in Banach spaces and adjoint operator of a compact operator is compact [13], we have from Lemma 3.1, (iii) that for any $\lambda \notin \sigma_p(\mathbf{B}^*)$, $(\mathbf{B}^* - \lambda I)^{-1}$ is compact with the point spectrum only. □

5. Eigenfunction expansions and little Hilbert spaces

In this section we describe the subspaces where Φ and Φ^* are closed, i.e., where there exist eigenfunction expansions of the elements.

5.1. **Operator B: Hilbert spaces $\tilde{L}^2_\rho$, $\tilde{H}^{2m}_\rho$, l^2_ρ and h^{2m}_ρ.** Let us first introduce some subspaces in L^2_ρ, where the complete eigenfunction subset Φ of operator $\mathbf{B}$ is closed. As usual, we define the linear subspace $\tilde{L}^2_\rho$ of eigenfunction expansions,

$$v \in \tilde{L}^2_\rho \quad \text{iff} \quad v = \sum c_\beta \psi_\beta \ \text{with convergence in} \ L^2_\rho, \tag{5.1}$$

as the closure of the subset of finite sums $\{\sum_{|\beta| \le K} c_\beta \psi_\beta, \ K \in \mathbb{N}\}$ in the L^2_ρ-norm. By the completeness-closure of Φ in $\tilde{L}^2_\rho$ and orthonormality (4.9), the expansion coefficients are

$$c_\beta = \langle v, \psi^*_\beta \rangle. \tag{5.2}$$

Since Φ is not orthonormal in L^2_ρ for $m > 1$, the strict inclusion takes place $\tilde{L}^2_\rho \subset L^2_\rho$, and the equality occurs in the self-adjoint case $m = 1$, $a = 1/4$ only. Actually, the difference $L^2_\rho \setminus \tilde{L}^2_\rho$ can measure a "defect" of non self-adjointness of operator $\mathbf{B}$ in L^2_ρ .

We next describe some properties of expansion coefficients for $v \in \tilde{L}^2_\rho$.

Proposition 5.1. *Let, for an arbitrarily small constant $\varepsilon > 0$, as $|\beta| \to \infty$,*

$$c_\beta = o(|\beta|^{|\beta|(\nu-\varepsilon)}), \quad \textit{where} \ \nu = (2-\alpha)/2\alpha > 0. \tag{5.3}$$

Then $v = \sum c_\beta \psi_\beta \in \tilde{L}^2_\rho$.

Proof. There holds

$$\int_{\mathbf{R}^N} \rho |v|^2 = \int \rho |\sum c_\beta \psi_\beta|^2 \equiv \sum_{(\beta,\gamma)} A_{\beta\gamma} c_\beta c_\gamma, \ \ A_{\beta\gamma} = \int \rho \psi_\beta \psi_\gamma. \tag{5.4}$$

Bearing in mind (3.6), it follows from standard kernel estimates [6] (cf. a sharp asymptotic estimate of the rescaled kernel in the right-hand side of (2.4)) that

$$|D^\beta f(y)| \le c^{|\beta|}(1+|y|)^{|\beta|(\alpha-1)} e^{-d|y|^\alpha} \quad \text{in} \ \mathbf{R}^N, \tag{5.5}$$

where c is independent of $|\beta|$. Therefore,

$$|A_{\beta\gamma}| \le \frac{c^{|\beta+\gamma|}}{\sqrt{\beta!\gamma!}} \int e^{-b|y|^\alpha} (1+|y|)^{|\beta+\gamma|(\alpha-1)}, \tag{5.6}$$

where $b = 2d - a > 0$ by the definition of the weight (1.13), (3.1). One can see that the right-hand side attains its minimal value for $|\beta| \sim |\gamma| = l \gg 1$ and then by Stirling's formula, omitting all the lower-order multiplier and keeping only those of the type given in (5.3),

$$\int_{\mathbf{R}^N} e^{-b|y|^\alpha} (1+|y|)^{|\beta+\gamma|(\alpha-1)} \sim \int_0^\infty z^{N-1} e^{-bz^\alpha} z^{2l(\alpha-1)} dz \sim \Gamma(\tfrac{2l(\alpha-1)}{\alpha}) \sim l^{2l(\alpha-1)/\alpha}.$$

This implies the estimate

$$|A_{\beta\gamma}| \sim (l!)^{-1} l^{2l(\alpha-1)/\alpha} \sim l^{-l} l^{2l(\alpha-1)/\alpha} = l^{l(\alpha-2)/\alpha}, \tag{5.7}$$

and hence (5.4) converges under assumption (5.3). $\square$

In fact, accurate using Stirling's formula shows that $\sum c_\beta \psi_\beta \in \tilde{L}^2_\rho$ if $c_\beta = O(\varepsilon^l |\beta|^{|\beta|\nu})$ with a sufficiently small $\varepsilon > 0$. Since estimates of the leading terms in (5.7) are sharp,

$$v = \sum c_\beta \psi_\beta \in \tilde{L}^2_\rho \implies c_\beta = o(|\beta|^{|\beta|(\nu+\varepsilon)}) \quad \text{for some } \varepsilon > 0. \tag{5.8}$$

By $\tilde{H}^{2m}_\rho \subset \tilde{L}^2_\rho$ we denote the dense linear subspace obtained as the closure in the norm of H^{2m}_ρ of the subset of eigenfunction expansions with coefficients satisfying (5.3). $\tilde{H}^{2m}_\rho$ with the scalar product of H^{2m}_ρ becomes a Hilbert space and can be considered as the domain of $\mathbf{B}$ in H^{2m}_ρ. There holds

$$\tilde{H}^{2m}_\rho \subseteq H^{2m}_\rho \cap \tilde{L}^2_\rho. \tag{5.9}$$

Note that (5.3) does not apply for $m = 1$ since then $\alpha = 2$ and hence $\nu = 0$. Actually, a natural optimal analogy of $\tilde{H}^{2m}_\rho$ for $m = 1$ is H^2_ρ, the domain of $\mathbf{B}$ in L^2_ρ.

We will need a subspace of $\tilde{L}^2_\rho$ introduced as a *little* Hilbert space l^2_ρ of functions $v = \sum c_\beta \psi_\beta \in \tilde{L}^2_\rho$ with coefficients satisfying

$$\sum |c_\beta|^2 < \infty, \tag{5.10}$$

where the scalar product and the induced norm are given by

$$(v, w)_0 = \sum c_\beta a_\beta, \text{ where } w = \sum a_\beta \psi_\beta \in l^2_\rho, \quad \text{and} \quad \|v\|^2_0 = (v, v)_0. \tag{5.11}$$

Obviously, l^2_ρ is isomorphic to the Hilbert space l^2 of sequences $\{c_\beta\}$ with the same inner product and hence

$$\Phi \quad \text{is orthonormal in } l^2_\rho. \tag{5.12}$$

It is worth mentioning that though Φ is not orthonormal in the big space L^2_ρ, estimates (5.6) show that after suitable orthogonalization according to sharp bounds (5.7), all the scalar products satisfy

$$|A_{\beta\gamma}| = |\langle \psi_\beta, \psi_\gamma \rangle_\rho| \ll 1 \quad \text{for all } |\beta| \gg |\gamma| = l \gg 1 \text{ or } 1 \ll |\beta| \ll |\gamma| \tag{5.13}$$

and actually are super-exponentially small for $l \gg 1$. This means that Gram's matrix $\Gamma = [A_{\beta\gamma}]$ of such a normalized Φ in L^2_ρ has a "diagonal dominance" in the sense that elements $|A_{\beta\gamma}| \ll 1$ if they stay sufficiently far from the main diagonal. Therefore, Φ is not "very much" non-orthogonal, and a standard Gram-Schmidt normalization of Φ performed by introducing the scalar product (5.11) of l^2_ρ seems to be quite natural.

We next define a little Sobolev space h^{2m}_ρ of functions $v \in l^2_\rho$ such that $\mathbf{B}v \in l^2_\rho$, i.e., $\sum |\lambda_\beta c_\beta|^2 < \infty$. The scalar product and the induced norm in h^{2m}_ρ are

$$(v, w)_1 = (v, w)_0 + (\mathbf{B}v, \mathbf{B}w)_0, \quad \|v\|^2_1 = (v, v)_1 \equiv \sum (1 + |\lambda_\beta|^2) |c_\beta|^2. \tag{5.14}$$

This norm is equivalent to the graph norm induced by the positive operator $(-\mathbf{B} + aI)$ with $a > 0$. Then h^{2m}_ρ is the domain of $\mathbf{B}$ in l^2_ρ. We also have a Sobolev embedding theorem,

$$h^{2m}_\rho \subset l^2_\rho \quad \text{compactly}, \tag{5.15}$$

which follows from the criterion of compactness in l^p, [13]. In the self-adjoint case $m = 1$, the little space l^2_ρ coincides with the big one,

$$l^2_\rho = L^2_\rho \quad \text{for } m = 1 \text{ if } a = \tfrac{1}{4} \text{ in (1.13).} \tag{5.16}$$

Then h^2_ρ is just the domain H^2_ρ of $\mathbf{B}$. If $a \neq 1/4$, then $\mathbf{B}$ is not self-adjoint in L^2_ρ and, in general, (5.16) is not true, $\tilde{L}^2_\rho \neq L^2_\rho$, even for $m = 1$.

Since the orthonormality of Φ is known to be of importance in the operator theory and applications, in some linear and nonlinear problems dealing with operators like $\mathbf{B}$, the little space l^2_ρ can play a special role in comparison with the big one L^2_ρ.

It follows from (5.11) that $\mathbf{B}$ is *self-adjoint* in l^2_ρ with the domain h^{2m}_ρ,

$$(\mathbf{B}v, w)_0 = (v, \mathbf{B}w)_0 \quad \text{for all } v, w \in h^{2m}_\rho. \tag{5.17}$$

Let us state other straightforward consequences (this list can be easily extended).

Proposition 5.2. *(i) l^2_ρ and is a dense subspace of $\tilde{L}^2_\rho$,*
(ii) $\Phi = \{\psi_\beta\}$ is complete and closed in l^2_ρ in the topology of L^2_ρ,
(iii) resolvent $(\mathbf{B} - \lambda I)^{-1}$ for $\lambda \notin \sigma(\mathbf{B})$ is compact in l^2_ρ, and
(iv) $\mathbf{B}$ is sectorial in l^2_ρ.

Proof. (i) Obviously, $l^2_\rho \subset \tilde{L}^2_\rho$ by Proposition 5.1. Concerning the density of l^2_ρ, we note that given a $v = \sum c_\beta \psi_\beta \in \tilde{L}^2_\rho$, the sequence of truncations $\{\sum_{|\beta| \leq K} c_\beta \psi_\beta, K \in \mathbb{N}\} \subset l^2_\rho$ converges to v in the topology of L^2_ρ as $K \to \infty$ by completeness and closure of $\{\psi_\beta\}$.

(ii) Since Φ is orthonormal in l^2_ρ, it follows that the only element orthogonal to $\{\psi_\beta\}$ is 0, and hence completeness of $\{\psi_\beta\}$ in l^2_ρ follows from the Riesz-Fischer theorem. It is closed as an orthonormal subset in a separable Hilbert space [13].

(iii) For any $v = \sum c_\beta \psi_\beta \in l^2_\rho$ from the unit ball T_1 in l^2_ρ with $\sum |c_\beta|^2 \leq 1$, $(\mathbf{B} - \lambda I)^{-1} v = \sum b_\beta \psi_\beta$, where

$$b_\beta = \frac{c_\beta}{\lambda_\beta - \lambda} = -\frac{c_\beta}{|\beta|/2m + \lambda} = -\frac{2mc_\beta}{|\beta|}[1 + O(\frac{1}{|\beta|})] \quad \text{for } |\beta| \gg 1. \tag{5.18}$$

Therefore, for any $\varepsilon > 0$, there exists $K = K(\varepsilon) > 0$ such that for any $v \in T_1$,

$$\textstyle\sum_{|\beta| \geq K} |b_\beta|^2 \leq 4m^2 K^{-2} \sum |c_\beta|^2 \leq 4m^2 K^{-2} < \varepsilon.$$

By the compactness criterion in l^2 [13], $(\mathbf{B} - \lambda I)^{-1}$ maps T_1 onto a compact subset in l^2_ρ.

(iv) Recall that $(\mathbf{B} - \lambda I)^{-1}$ is a meromorphic function having a pole $\sim 1/\lambda$ as $\lambda \to 0$ since $\lambda_0 = 0$ has multiplicity one [8]. We then need an extra estimate on the resolvent which is easy to get in l^2_ρ (it is not easy at all in the big space L^2_ρ). In the sector $\Phi_\theta = \{\lambda \in \mathbb{C} : \lambda \neq 0, |\arg \lambda| < \pi/2 + \theta\}$ with a $\theta \in (0, \pi/2)$, for any $v = \sum c_\beta \psi_\beta \in l^2_\rho$, we apply (5.18) by using that $1/|\lambda_\beta - \lambda| \leq 1/|\lambda| \sin\theta$ in Φ_θ to get

$$\|(\mathbf{B} - \lambda I)^{-1} v\|_0 = \left(\sum |c_\beta|^2 \frac{1}{|\lambda_\beta - \lambda|^2} \right)^{1/2} \leq \frac{1}{\sin\theta} \frac{1}{|\lambda|} \|v\|_0.$$

Since $\mathbf{B}$ is closed and densely defined, it is a sectorial operator in l^2_ρ, see [7]. $\square$

5.2. **Adjoint operator $\mathbf{B}^*$: Hilbert spaces $\tilde{L}^2_{\rho^*}$, $\tilde{H}^{2m}_{\rho^*}$, $l^2_{\rho^*}$ and $h^{2m}_{\rho^*}$.** Similarly, for the adjoint operator $\mathbf{B}^*$, we define subspace $\tilde{L}^2_{\rho^*} \subseteq L^2_{\rho^*}$, where the eigenfunction subset Φ^* is closed (cf. (5.1)), i.e., $v = \sum c_\beta \psi^*_\beta$ with

$$c_\beta = \langle v, \psi_\beta \rangle. \tag{5.19}$$

Proposition 5.3. *If $v = \sum c_\beta \psi^*_\beta \in \tilde{L}^2_{\rho^*}$, then for arbitrarily small $\varepsilon > 0$,*

$$c_\beta = o(|\beta|^{-|\beta|(\nu-\varepsilon)}) \quad \textit{for } |\beta| \gg 1, \qquad \nu = (2-\alpha)/2\alpha. \tag{5.20}$$

Proof. Similar to (5.4), we have

$$\int \rho^* |v|^2 = \textstyle\sum_{(\beta,\gamma)} A^*_{\beta\gamma} c_\beta c_\gamma, \quad A^*_{\beta\gamma} = \int \rho^* \psi^*_\beta \psi^*_\gamma, \tag{5.21}$$

where by (4.4) and Stirling's formula we estimate the coefficients for $|\beta| = |\gamma| = l$,

$$|A^*_{\beta\gamma}| \sim \frac{c^{|\beta|}}{\sqrt{\beta!\gamma!}} \int e^{-a|y|^\alpha} (1+|y|)^{|\beta+\gamma|} \sim \frac{1}{l!} \Gamma(\tfrac{2l}{\alpha}) \sim l^{-l} l^{2l/\alpha} = l^{l(2-\alpha)/\alpha}. \tag{5.22}$$

Hence, (5.20) is a necessary condition for convergence of series (5.21). □

One can see from such estimates that

$$c_\beta = o(|\beta|^{-|\beta|(\nu+\varepsilon)}) \implies \sum c_\beta \psi^*_\beta \in \tilde{L}^2_{\rho^*}. \tag{5.23}$$

Remark. Conditions (5.3) and (5.23) on the expansion coefficients in l^2_ρ and $l^2_{\rho^*}$ are different and do not exhibit a natural symmetry unlike the self-adjoint case $m = 1$. The symmetry can be restored by introducing normalization multipliers, $(\beta!)^{-(\alpha-1)/\alpha}$ and $(\beta!)^{-1/\alpha}$, in (3.6) and (4.4) respectively. For $m = 1$, where $\alpha = 2$, both are equal to $(\beta!)^{-1/2}$, which we continue to use for any $m > 1$ in our analysis below.

By $\tilde{H}^{2m}_{\rho^*}$ we denote the closure in the norm of $H^{2m}_{\rho^*}$ of the linear subspace of eigenfunction expansions with coefficients satisfying (5.20) for some $\varepsilon > 0$. Being equipped with the scalar product (3.3) with $\rho \mapsto \rho^*$, $\tilde{H}^{2m}_{\rho^*}$ is a Hilbert space becoming the domain of $\mathbf{B}^*$ in $H^{2m}_{\rho^*}$. We have

$$\tilde{H}^{2m}_{\rho^*} \subseteq H^{2m}_{\rho^*} \cap \tilde{L}^2_{\rho^*}. \tag{5.24}$$

In view of the fast decay (5.23) of the coefficients, similar to l^2_ρ, we introduce the adjoint little Hilbert space $l^2_{\rho^*}$ of eigenfunction expansions $v = \sum c_\beta \psi^*_\beta \in \tilde{L}^2_{\rho^*}$ with the scalar product $(\cdot,\cdot)_{0*}$ and the norm $\|\cdot\|_{0*}$ defined as in (5.11). As the domain of $\mathbf{B}^*$ in $l^2_{\rho^*}$, we introduce the corresponding little Sobolev space $h^{2m}_{\rho^*}$ compactly embedded into $l^2_{\rho^*}$, and by $(\cdot,\cdot)_{1*}$ and $\|\cdot\|_{1*}$ we denote the scalar product and the induced norm.

Then $\mathbf{B}^*$ is self-adjoint in $l^2_{\rho^*}$, and, obviously, $\tilde{L}^2_{\rho^*}$ and $\tilde{H}^{2m}_{\rho^*}$ are dense subspaces of $l^2_{\rho^*}$.

6. Classification of multiple zeros

We now return to the perturbed equation (1.7), where the exponential perturbation $\mathbf{C}(\tau)$ includes operators of $2m$-th order. Therefore, application of classical estimates on semigroups generated by sectorial operators [7], [11], based on fractional powers of

operators, are not straightforward. We also note that known detailed study of similar asymptotics of multiple zeros for the second-order parabolic equations [12], [1], [4], [3] often essentially uses estimates and other features of the self-adjoint rescaled operator $\mathbf{B}^*$ ($m = 1$), and do not apply to the non self-adjoint case $m > 1$, though some abstract general ideas and results from [12] and [1] are quite effective and will be used later on.

Assuming that $u(\tau) \in \tilde{L}^2_{\rho^*}$ for $\tau > 0$, we will use the eigenfunction expansion of such solutions of equation (1.7)

$$u(\tau) = \sum c_\beta(\tau)\psi^*_\beta \quad \text{with coefficients } c_\beta(\tau) = \langle u(\tau), \psi_\beta \rangle, \tag{6.1}$$

see (5.19). We next impose extra smoothness conditions on the solution and on the coefficients of the equation. We suppose that $u(\tau)$ is uniformly bounded in $H^{2m}_{\rho^*}$,

$$\|u(\tau)\|_{2m,*} \le C \quad \text{for all } \tau > 0, \tag{6.2}$$

a natural *a priori* bound in the parabolic theory, [6, 7]. Assuming that $u(\tau) \in \tilde{L}^2_\rho$, by (5.20) we may suppose that

$$c_\beta(\tau) = o(|\beta|^{-|\beta|(\nu-\varepsilon)}) \to 0 \quad \text{as } |\beta| \to \infty \quad \text{uniformly in } \tau \in [1, \infty). \tag{6.3}$$

The rate of decay (5.23) is sufficient for performing manipulations various series.

Substituting (6.1) into equation (1.7) and multiplying in L^2 by the adjoint eigenfunction ψ_β, we obtain the following system on the coefficients:

$$\dot{c}_\beta = \lambda_\beta c_\beta + J_\beta(\tau), \quad \text{where } J_\beta(\tau) \equiv \langle \mathbf{C}(\tau) \textstyle\sum_{(\gamma)} c_\gamma \psi^*_\gamma, \psi_\beta \rangle \quad \text{for any } \beta. \tag{6.4}$$

Using (1.9) and integrating by parts yield the perturbation consisting of two terms

$$J_\beta(\tau) = J_{\beta 1}(\tau) + J_{\beta 2}(\tau), \quad \text{where}$$

$$J_{\beta 1}(\tau) = \Big\langle \sum_{|\mu|=2m} R_\mu(\tau) D^\mu \sum_{(\gamma)} c_\gamma \psi^*_\gamma, \psi_\beta \Big\rangle \equiv \sum_{|\mu|=2m} \sum_{(\gamma)} c_\gamma g_{\mu\gamma\beta}(\tau), \tag{6.5}$$

$$J_{\beta 2}(\tau) = \Big\langle \sum_{|\mu|<2m} e^{-(2m-|\mu|)\tau/2m} a_\mu(\tau) D^\mu \sum_{(\gamma)} c_\gamma \psi^*_\gamma, \psi_\beta \Big\rangle$$

$$\equiv \sum_{|\mu|<2m} e^{-(2m-|\mu|)\tau/2m} \sum_{(\gamma)} c_\gamma h_{\mu\gamma\beta}(\tau), \tag{6.6}$$

$$g_{\mu\gamma\beta}(\tau) = \langle R_\mu(\tau) D^\mu \psi^*_\gamma, \psi_\beta \rangle, \quad h_{\mu\gamma\beta}(\tau) = \langle a_\mu(\tau) D^\mu \psi^*_\gamma, \psi_\beta \rangle. \tag{6.7}$$

By (1.10), we set

$$g_{\mu\gamma\beta}(\tau) = e^{-\tau/2}\, \tilde{g}_{\mu\gamma\beta}(\tau).$$

Since the exponential estimates (1.10) and (1.11) hold on compact subsets only, in the higher-order perturbation term (6.5) one needs to estimate the integrals over $\{|y| \ge r\}$ with $r = r(\tau) \gg 1$, where we just assume that $R_\mu(\tau)$ are uniformly bounded, and hence, similar to (5.22) in the essential "diagonal" cases with $|\beta| \sim |\gamma| \sim l \gg 1$,

$$\Big| \int_{|y| \ge r} R_\mu(\tau) D^\mu \psi^*_\gamma \psi_\beta \Big| \sim (l!)^{-1} \int_r^\infty e^{-dz^\alpha} z^{l\alpha+N-1} dz \sim e^{-dr^\alpha/2} \tag{6.8}$$

(as usual, we keep the leading multiplier only which is sufficient for necessary rough estimates). Therefore, choosing $r(\tau) \sim \tau^{1/\alpha}$ for $\tau \gg 1$, we obtain the same exponential factors with, possibly, some extra multipliers with not more than algebraic growth as $\tau \to \infty$ which will be omitted.

As the necessary hypothesis on the coefficients $a_\beta(x,t)$ of the parabolic equation (1.1), we assume that the multipliers $\tilde{g}_{\mu\gamma\beta}(\tau)$ and $h_{\mu\gamma\beta}(\tau)$ are bounded and, under assumption (6.3), the corresponding series in (6.5) and (6.6) converges sufficiently fast. The regularity hypotheses can be weakened but are a convenient restriction for further asymptotic analysis.

Summing up the above manipulations, we arrive at the following infinite-dimensional dynamical system on the expansion coefficients:

$$\dot{c}_\beta = \lambda_\beta c_\beta + \textstyle\sum_{(\mu,\gamma)} e^{-\nu_\mu \tau} j_{\mu\gamma\beta}(\tau) c_\gamma, \quad \text{where} \tag{6.9}$$

$$\nu_\mu = (2m - |\mu|)/2m \quad \text{for } |\mu| < 2m \quad \text{and} \quad \nu_\mu = 1/2 \text{ for } |\mu| = 2m, \tag{6.10}$$

and $j_{\mu\gamma\beta}(\tau)$ are bounded coefficients related to $\tilde{g}_{\mu\gamma\beta}$ and $h_{\mu\gamma\beta}$. The series on the right-hand side of (6.9) converges sufficiently fast.

6.1. **Multiple zeros for the unperturbed equation.** A complete classification of the multiple zeros is straightforward for the unperturbed equation (1.7) with the null operator $\mathbf{C} = 0$. Then (6.4) takes the diagonal form $\dot{c}_\beta = \lambda_\beta c_\beta$ and hence $c_\beta(\tau) = C_\beta e^{\lambda_\beta \tau}$ with $C_\beta = c_\beta(0)$ for any β. Therefore, this linear homogeneous parabolic equation admits different types of formation of multiple zeros given by the countable subset of patterns

$$u_\beta(y,\tau) = e^{\lambda_\beta \tau} \psi^*_\beta(y), \quad |\beta| \geq 1. \tag{6.11}$$

The first pattern with $\beta = 0$ and $\lambda_0 = 0$ is excluded since $\psi^*_0 \equiv 1$ and the corresponding pattern $u_0(y,\tau) \equiv 1$ does not vanish. Using (4.4), denote

$$P_0(y) = (l!)^{-1/2} \textstyle\sum_{|\beta|=l} C_\beta y^\beta \not\equiv 0$$

the homogeneous polynomial of l-th order. Then similar to (4.4),

$$\varphi^*_l(y) = P_0(y) + \textstyle\sum_{j=1}^{[l/2m]} \frac{1}{j!} (-\mathbf{B}_0)^j P_0(y). \tag{6.12}$$

Thus, a general structure of zero surfaces of l-th order multiple zero is given by the nodal set of an eigenfunction $\varphi^*_l(y)$, i.e., by a nontrivial linear combination

$$\varphi^*_l(y) = \textstyle\sum_{|\beta|=l} C_\beta \psi^*_\beta(y) \not\equiv 0. \tag{6.13}$$

Namely, if $C_\beta = 0$ for any $|\beta| < l$ and there exists a $C_\beta \neq 0$ for a $|\beta| = l$, then the asymptotic behaviour of the corresponding solution is as follows:

$$u(y,\tau) = e^{-l\tau/2m} [\varphi^*_l(y) + O(e^{-\tau/2m})] \quad \text{as } \tau \to \infty. \tag{6.14}$$

If such a finite l does not exist, then the solution is trivial, $u \equiv 0$. This provides us with the first backward uniqueness result to be extended later on to more general equations. It is elementary for such solutions $u(y,\tau)$ which are analytic in y.

6.2. **Perturbed equation.** We will perform a similar analysis of the perturbed dynamical system (6.9).

STEP 1. We begin with the first equation for $c_0(\tau)$ with $\beta = 0$ and $\lambda_0 = 0$, where by (6.3), the right-hand side can be estimated as follows:

$$\dot{c}_0 = \textstyle\sum_{(\mu,\gamma)} e^{-\nu_\mu \tau} j_{\mu\gamma 0} c_\gamma = O(e^{-\tau/2m}) \quad \text{for } \tau \gg 1, \tag{6.15}$$

meaning that $|\dot{c}_0(\tau)| \le A e^{-\tau/2m}$ for all $\tau \ge 0$, where $A > 0$ is a constant. Hence, there exists a finite limit $C_0 = c_0(\infty)$, and integrating equation over (τ, ∞) yields

$$c_0(\tau) = C_0 + O(e^{-\tau/2m}). \tag{6.16}$$

Let us estimate coefficients $c_\beta(\tau)$ with $|\beta| \ge 1$. Writing down the equations in the form

$$(c_\beta(\tau) e^{-\lambda_\beta \tau})' = e^{-\lambda_\beta \tau} \textstyle\sum_{(\mu,\gamma)} e^{-\nu_\mu \tau} j_{\mu\gamma\beta} c_\gamma = O(e^{-(\lambda_\beta + 1/2m)\tau}) \tag{6.17}$$

and integrating over $(0, \tau)$ with $\tau \gg 1$ yields

$$c_\beta(\tau) = c_\beta(0) e^{\lambda_\beta \tau} + e^{\lambda_\beta \tau} \int_0^\tau O(e^{-(\lambda_\beta + 1/2m)s}) ds = O(\tau e^{-\tau/2m}), \tag{6.18}$$

where an extra power multiplier τ is taken into account in the resonance case $\lambda_\beta = -1/2m$, i.e., for $|\beta| = 1$. Here and later on, in similar estimates we omit the dependence of the coefficients on $|\beta|$ which is covered by assumptions of fast convergence of the series involved. It is important that under the above hypotheses on $j_{\mu\gamma\beta}(\tau)$ and (6.3), estimates (6.17) and (6.18) are uniform in β.

Thus, in view of the above hypotheses on the solution, if $C_0 \ne 0$, then one obtains

$$u(y, \tau) = C_0 + o(1), \tag{6.19}$$

i.e., as in the unperturbed case, solutions do not exhibit zero formation as $\tau \to \infty$ on any compact subset in y. The estimate $o(1) \sim O(\tau e^{-\tau/2m})$ obtained via the above manipulations remains valid up to an extra not more than algebraically growing multipliers via the perturbation terms in (6.9).

Hence, we assume that $C_0 = 0$. Then by (6.18)

$$c_\beta(\tau) = O(\tau e^{-\tau/2m}) \quad \text{uniformly in } |\beta| \ge 0. \tag{6.20}$$

Substituting these estimates into (6.15) yields a refined estimate on the first coefficient

$$\dot{c}_0 = O(\tau e^{-\tau/m}) \implies c_0(\tau) = O(\tau e^{-\tau/m}). \tag{6.21}$$

STEP 2. Consider next equations with $|\beta| = 1$, $\lambda_\beta = -1/2m$, where we use estimates (6.20) to get that $\dot{c}_\beta = -\frac{1}{2m} c_\beta + \sum_{(\mu,\gamma)} e^{-\nu_\mu \tau} j_{\mu\gamma\beta} c_\gamma = -\frac{1}{2m} c_\beta + O(\tau e^{-\tau/m})$. Multiplying by $e^{\tau/2m}$,

$$(c_\beta(\tau) e^{\tau/2m})' = O(\tau e^{-\tau/2m}), \tag{6.22}$$

we have that there exists a finite limit $c_\beta(\tau) e^{\tau/2m} \to C_\beta$ as $\tau \to \infty$. Integrating (6.22) over (τ, ∞), we deduce

$$c_\beta(\tau) = C_\beta e^{-\tau/2m} + O(\tau e^{-\tau/m}), \quad |\beta| = 1. \tag{6.23}$$

If $C_\beta \neq 0$ for some β with $|\beta| = 1$, then using (6.20), from equations with $|\beta| \geq 2$ we estimate the coefficients as follows:

$$(c_\beta e^{-\lambda_\beta \tau})' = O(\tau e^{-(\lambda_\beta + 1/m)\tau}), \tag{6.24}$$

and integrating over $(0, \tau)$, we get that

$$c_\beta(\tau) = O(\tau^2 e^{-\tau/m}), \quad |\beta| \geq 2. \tag{6.25}$$

It follows from (6.21), (6.25) and (6.23) that for all $|\beta| \neq 1$ and those $|\beta| = 1$ with $C_\beta = 0$ the expansion coefficients satisfy

$$c_\beta(\tau) = O(\tau^2 e^{-\tau/m}) \tag{6.26}$$

uniformly in $|\beta|$. Hence, in this case (6.23) implies the asymptotic behaviour

$$u(y, \tau) = e^{-\tau/2m}[\varphi_1^*(y) + o(1)], \tag{6.27}$$

with the eigenfunction φ_1^* given by (6.13) and $o(1) \sim O(\tau^2 e^{-\tau/2m})$ with, possibly, an extra algebraic factor.

STEP l. We iterate the dynamical system $l - 1$ times assuming that the limits

$$c_\beta(\tau) e^{-\lambda_\beta \tau/2m} \to C_\beta \quad \text{as } \tau \to \infty \tag{6.28}$$

are trivial, $C_\beta = 0$, for all $|\beta| = 0, 1, ..., l - 1$, and there exists a first β, $|\beta| = l$, such that $C_\beta \neq 0$ (as above, existence of such limits follows from the convergence of integrals). Then, similarly to the previous analysis, we derive that

$$c_\beta(\tau) = C_\beta e^{-l\tau/2m} + O(\tau^l e^{-(l+1)\tau/2m}) \quad \text{for } |\beta| = l \text{ and } C_\beta \neq 0, \quad \text{and} \tag{6.29}$$

$$c_\beta(\tau) = O(\tau^{l+1} e^{-(l+1)\tau/2m}) \quad \text{for } |\beta| \neq l \text{ or } |\beta| = l \text{ and } C_\beta = 0. \tag{6.30}$$

Note again that (6.30) are uniform in $|\beta| \gg 1$. Therefore, the corresponding multiple zero pattern has the form

$$u(y, \tau) = e^{-l\tau/2m}[\varphi_l^*(y) + o(1)], \tag{6.31}$$

with the eigenfunction (6.13) and $o(1) \sim O(\tau^{l+1} e^{-\tau/2m})$ with, possibly, an extra factor of the algebraic growth. This completes the classification of finite order multiple zeros.

We next need to prove that infinite order zeros exist for trivial solutions $u \equiv 0$ only, i.e., dynamical system (6.9) does not admit nontrivial solutions with a super-exponential decay rate as $\tau \to \infty$. In the abstract form, for linear equations in Hilbert spaces, such results are well established in the mathematical literature, see [1], [3], [4], [12] and earlier references on Agmon's and Ogawa's results therein. Some of the approaches essentially rely on the symmetry of the unperturbed operator $\mathbf{B}^*$ (cf. Section 5 in [3]) and therefore cannot be applied here. We follow the lines of the analysis in [1], Appendix, which is formulated for self-adjoint operators but admits a natural extension to general operators $\mathbf{B}^*$ with real spectrum bounded from above.

Proposition 6.1. *Assume that $u(\tau) \in \tilde{H}^{2m}_{\rho^*}$ is such that*

$$\|u(\tau)\|_{0*} = o(e^{-K\tau}) \quad \text{as } \tau \to \infty, \tag{6.32}$$

where K can be arbitrarily large constant. Then $u \equiv 0$.

Proof. We follow the lines and basic notations of the analysis in [1], pp. 434-437. We have that the operator $\mathbf{A} = -\mathbf{B}^* : E_1 = h^{2m}_{\rho^*} \to l^2_{\rho^*} = E$ has a nonnegative discrete spectrum $\sigma(\mathbf{A}) = \{\tilde{\lambda}_\beta = |\beta|/2m\}$ with the constant spectral half-gaps denoted by $\delta_k \equiv \delta = 1/4m$. Then we set $\gamma_k = (2k+1)/4m$. The spectral projections of $\mathbf{A}$ denoted by $P_k = P(\gamma_k)$ for $k = 1, 2, ...$ are given by

$$P_k u = \sum_{|\beta| \le k} \langle u, \psi_\beta \rangle \psi^*_\beta, \quad \text{and} \quad Q_k u = (I - P_k)u = \sum_{|\beta| > k} \langle u, \psi_\beta \rangle \psi^*_\beta.$$

We next check conditions (B1) and (B2) in [1], p. 435. Under given assumptions on the smoothness of coefficients of the parabolic operator, the perturbation operator $\mathbf{C}(\tau)$ in (1.7) is Hölder continuous with respect to the operator norm on $\mathcal{L}(E_1, E)$. Since $\mathbf{C}(\tau)$ is a differential $2m$-th order operator with smooth exponentially small coefficients given in (1.9) and (1.10), we have that $\mathbf{C}(\tau) : E_1 \to E$ is a bounded operator,

$$\|\mathbf{C}(\tau)u\|_{0*} \le \|\mathbf{C}(\tau)\| \, \|u\|_{1*}, \quad \text{where } \|\mathbf{C}(\tau)\| = O(e^{-\tau/2m}) \text{ for } \tau \gg 1.$$

By shifting the origin in time, we may assume that

$$M = \sup_{\tau \ge 0} \|\mathbf{C}(\tau)\| < \delta/2,$$

i.e., $\|\mathbf{C}(\tau)\|$ is small in comparison to the gaps in the spectrum of $\mathbf{A}$. Denote by $u(\tau) = S(\tau)u_0$ the unique sufficiently smooth solution of (1.7) with initial data $u_0 \in E$, see [11] and [7]. We next introduce the subspaces

$$V_k = \{u_0 \in E_1 : \quad e^{\gamma_k \tau} S(\tau) u_0 \to 0 \text{ as } \tau \to \infty\},$$

where $V_{k+1} \subset V_k$ for any $k \ge 1$.

We now apply Lemma 5 in [1] which is proved in similar lines without using specific self-adjoint properties of the unperturbed operator $\mathbf{A} = -\mathbf{B}^*$. This part is based on the analysis of the integral equation

$$u(\tau) = e^{\mathbf{B}^* \tau} Q_k u_0 + \int_0^\tau e^{\mathbf{B}^*(\tau - s)} Q_k \mathbf{C}(s) u(s) ds - \int_\tau^\infty e^{\mathbf{B}^*(\tau - s)} P_k \mathbf{C}(s) u(s) ds.$$

Setting, $v(\tau) = e^{\gamma_k \tau} u(\tau)$ gives the integral equation

$$v(\tau) = e^{(\gamma_k + \mathbf{B}^*)\tau} Q_k u_0 + \int_0^\infty K(\tau - s)\mathbf{C}(s)v(s)ds \equiv q(\tau) + Lv(\tau), \tag{6.33}$$

with the kernel

$$K(\nu) = \{Q_k e^{(\gamma_k + \mathbf{B}^*)\nu} \text{ if } \nu > 0 \text{ and } -P_k e^{(\gamma_k + \mathbf{B}^*)\nu} \text{ if } \nu < 0\}.$$

It follows that for $\nu > 0$ and any $w \in h^{2m}_{\rho^*}$, $\|K(\nu)w\|^2_{0*} = \sum_{|\beta| > k} e^{2(\gamma_k + \lambda_\beta)\nu} |c_\beta|^2 \le e^{-2\delta\nu} \|w\|^2_{0*} \le e^{-2\delta\nu} \|w\|^2_{1*}$, where $A^*_{\beta\gamma}$ is as estimated below (5.21), so that in operator norms on $\mathcal{L}(E)$ and $\mathcal{L}(E_1, E)$, $\|K(\nu)\| \le e^{-\delta\nu}$ for $\nu > 0$. By a similar estimate for $\nu < 0$, we conclude that the kernel is exponentially decaying,

$$\|K(\nu)\| \le e^{-\delta|\nu|} \quad \text{for } \nu \in \mathbf{R}. \tag{6.34}$$

Equation (6.33) can be solved by Banach's Contraction Principle in the space

$$F = \{v \in C([0, \infty); E_1) : \quad v(\tau) \to 0 \quad \text{as } \tau \to \infty\}.$$

We have

$$\|Lv(\tau)\|_{0*} \le \int_0^\infty \|K(t-s)\|\,\|\mathbf{C}(s)\|\,\|v(s)\|_{1*}ds \le M\left(\int_{-\infty}^\infty \|K(\nu)\|d\nu\right)\sup_{s\ge 0}\|v(s)\|_{1*},$$

and hence by (6.34)

$$\|L\| \le M\int_{-\infty}^\infty \|K(\nu)\|d\nu \le 2M/\delta < 1.$$

Therefore, the solution of (6.33) is given by the converging series $v(\tau) = \sum_{j=0}^\infty L^j q(\tau)$. Denote $q_0 = Q_k u_0 \in R(Q_k)$. Then the mapping T_k defined by $v(0) = T_k q_0$ is bounded,

$$\|T_k\| \le \textstyle\sum_{j=0}^\infty (2M/\delta)^j = (1-2M/\delta)^{-1} < \infty. \tag{6.35}$$

Since $Q_k \circ T_k = I$ on $R(Q_k)$ and $T_k \circ Q_k = I$ on V_k, it follows that $Q_k : V_k \to R(Q_k)$ is an isomorphism; see [1], p. 435. Hence, if $u_0 \in V_k$ for all $k \ge 1$, then $u_0 = T_k \circ Q_k u_0$ and

$$\|u_0\|_{0*} = \|T_k \circ Q_k u_0\|_{0*} \le (1-2M/\delta)^{-1}\|Q_k u_0\|_{0*},$$

where $Q_k u_0 = \sum_{|\beta|>k}\langle u_0, \psi_\beta\rangle\psi_\beta^* \to 0$ as $k \to \infty$. Therefore, $u_0 = 0$. This completes the proof of Proposition 6.1. □

Thus, we arrive at the following classification of multiple zeros of solutions to (1.1).

Theorem 6.1. *Let, under given regularity assumptions* (6.2), (6.3), *the rescaled solution* $u(\cdot,\tau) \in \tilde{H}^{2m}_{\rho^*}$ *of* (1.1) *create a multiple zero at* $(0,0)$. *Then there exists a finite* $l \ge 1$ *such that* (6.31) *holds, where* φ_l *is an eigenfunction* (6.13) *of* $\mathbf{B}^*$ *corresponding to the eigenvalue* $-l/2m$.

Next, we need to interpret the above asymptotic result by using the standard time-independent parabolic rescaling

$$u_\varepsilon(y,s) = \varepsilon^{-l}u(y\varepsilon, s\varepsilon^{2m}), \quad \text{with an arbitrary parameter } \varepsilon > 0, \tag{6.36}$$

where $s < 0$ is the new time variable. Then u_ε satisfies the perturbed equation (cf. (1.7))

$$u_s = \mathbf{B}_0 u + \mathbf{C}(\varepsilon)v, \tag{6.37}$$

with an asymptotically small perturbing operator

$$\mathbf{C}(\varepsilon) = \textstyle\sum_{|\beta|=2m}[a_\beta(y\varepsilon, s\varepsilon^{2m}) - A_\beta]D_y^\beta + \sum_{|\beta|<2m}\varepsilon^{2m-|\beta|}a_\beta(y\varepsilon, s\varepsilon^{2m})D_y^\beta. \tag{6.38}$$

By Theorem 6.1 we arrive at the following straightforward consequence.

Corollary 6.1. *With an* l *as in Theorem* 6.1, $\{u_\varepsilon\}_{\varepsilon>0}$ *is a compact subset, and uniformly on compact subsets from* $\mathbf{R}^N \times (-\infty, 0]$, *there holds*

$$u_\varepsilon(y,s) \to W(y,s) = (-s)^{l/2m}\varphi_l^*(y/(-s)^{1/2m}) \quad \text{as } \varepsilon \to 0^+. \tag{6.39}$$

Note that by (6.12), there exists a finite limit

$$W(y,0^-) = P_0(y) \equiv (l!)^{-1/2}\textstyle\sum_{|\beta|=l} C_\beta y^\beta \not\equiv 0. \tag{6.40}$$

6.3. **Instantaneous collapse of multiple zeros.** We now consider the evolution of the above solutions for $t > 0$ describing collapse of multiple zeros. For the one-dimensional second-order parabolic equations such extension was also performed by Sturm [18]. In was shown that, due to the established asymptotic behaviour as $t \to 0^+$ driven by the adjoint polynomials, zero curves disappear at $t = 0$ in each of such collapse. This led Sturm to state his remarkable First Theorem saying that the number of zeros of solutions does not increase with time; see p. 431 in [18].

We briefly describe the collapse phenomenon of multiple zeros for the higher-order equations under consideration. We apply a time-evolution description of such transition phenomenon from $\{t < 0\}$ to $\{t > 0\}$. Consider a general multiple zero pattern (6.31). Bearing in mind the Sturm variable (1.4), we have that

$$u(x,t) = (-t)^{l/2m}\varphi_l^*(x/(-t)^{1/2m}) + ..., \tag{6.41}$$

where we omit higher-order terms. Since $\varphi_l^*(y)$ is a polynomial of order l, by (4.4)

$$\varphi_l^*(y) = (l!)^{-1/2}\textstyle\sum_{|\beta|=l} C_\beta y^\beta + ... \equiv P_0(y) + ... \quad \text{as } y \to \infty, \tag{6.42}$$

where $P_0(y)$ denotes a nontrivial homogeneous polynomial of order l. Therefore, passing to the limit $t \to 0^-$ in (6.41) and using (6.42), we observe that in the leading term the time dependent multipliers cancel each other. It follows from Corollary 6.1 that there exists a finite limit

$$u(x,0^-) = (l!)^{-1/2}\textstyle\sum_{|\beta|=l} C_\beta x^\beta + ... \equiv P_0(x) + ... \quad \text{for small } x. \tag{6.43}$$

For convenience, we now perform a formal evolution analysis to be justified later on. Introducing the *forward* independent variables

$$y = x/t^{1/2m}, \quad \tau = \ln t,$$

we arrive at an exponentially perturbed equation of the form

$$u_\tau = \tilde{\mathbf{B}}u + \mathbf{C}(\tau)u, \quad \text{where } \tilde{\mathbf{B}} = \mathbf{B} - \tfrac{N}{2m}I, \tag{6.44}$$

with (6.43) as the initial data. The limit $t \to 0^+$ means $\tau \to -\infty$. As usual, this asymptotic problem with *a priori* prescribed initial data is easier than the above evolution one. We then obtain that the asymptotic behaviour as $\tau \to -\infty$ is given by adjoint polynomials associated with the operator $\tilde{\mathbf{B}}$,

$$u(y,\tau) = e^{l\tau/2m}\Phi_l(y) + ... \tag{6.45}$$

Similar to (6.12), we obtain the following representation of such polynomials:

$$\Phi_l(y) = P_0(y) + \textstyle\sum_{j=1}^{[l/2m]} \frac{1}{j!}\mathbf{B}_0^j P_0(y). \tag{6.46}$$

Note that $\Phi_l \notin L^2_\rho$ are not eigenfunctions of $\tilde{\mathbf{B}}$. Therefore, in the original variables,

$$u(x,t) = t^{l/2m}\Phi_l(x/t^{1/2m}) + ..., \tag{6.47}$$

and hence taking into account the leading higher-order terms in polynomial (6.46), we see that $u(x,0^+) = P_0(x) + ...$ coinciding with (6.43).

The generating formulas of polynomials (6.12) (for $t < 0$) and (6.46) (for $t > 0$) describe all possible exchanges of zero surfaces at the focusing time $t = 0$ of multiple zeros. Combining both expansions (6.31) and (6.47), we have that if a multiple zero of $u(x,t)$ occurs at the origin $(0,0)$, then there exists a finite $l \geq 1$ such that as $\varepsilon \to +0$,

$$\varepsilon^{-l/2m} u(y\varepsilon^{1/2m}, -\varepsilon) \to \varphi_l^*(y) \quad \text{and} \quad \varepsilon^{-l/2m} u(y\varepsilon^{1/2m}, \varepsilon) \to \Phi_l(y) \tag{6.48}$$

uniformly on compact subsets.

We now apply the rescaling argument based on the transformation (6.36) leading to the perturbed equations (6.37). Since the rescaling makes sense for both $s < 0$ and $s > 0$ and the equation and the asymptotically small perturbation operator in (6.38) are of the same structure, we arrive at the following result (for the second-order case $m = 1$, see [3]).

Corollary 6.2. *Under the assumptions of Corollary* 6.1, (6.39) *holds uniformly on compact subsets in* $\mathbf{R}^N \times [0, \infty)$.

We thus obtain that the function $W(y,s)$ in (6.39) is a polynomial solution of the linear homogeneous parabolic equation

$$W_s = \mathbf{B}_0 W \equiv \textstyle\sum_{|\beta|=2m} A_\beta D_y^\beta W \quad \text{in } \mathbf{R}^N \times \mathbf{R}. \tag{6.49}$$

Hence,

$$\Phi_l(y) = (-1)^{l/2m} \varphi_l^*(y/(-1)^{1/2m}), \tag{6.50}$$

which is seen from (6.12) and (6.46).

7. Unique continuation theorem

The first unique continuation theorem is a consequence on the above result establishing that a solution from the existence-uniqueness class of the parabolic equation (1.1) with sufficiently smooth coefficients cannot generate a multiple zero of infinite order unless $u \equiv 0$.

Theorem 7.1. *Let, under given hypotheses on the coefficients, the solution* $u(\cdot,t) \in \tilde{H}^{2m}_{\rho^*}$ *of* (1.1) *satisfy*

$$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon^k} \int_{|x|<\varepsilon} |u(x,0)| dx = 0 \quad \textit{for any } k > 0. \tag{7.1}$$

Then $u = 0$ *in* $\mathbf{R}^N \times (-1,1)$.

Proof. In view of (6.43), the integral condition (7.1) implies that the solution $u(x,t)$ has a zero in infinite multiplicity at the origin $(0,0)$, and hence $u = 0$ by Proposition 6.1 □

Obviously, once we have achieved the optimal classification of multiple zeros (the microstructure of the PDE), some backward uniqueness results are straightforward. Actually, one can characterize a variety of such optimal backward uniqueness approaches as follows:

(7.2) if at $(0,0)$ a solution u violates (6.31) (or (6.43)) for any $l \in \mathbb{N}$, then $u \equiv 0$.

The results apply to systems of $2m$-th order linear parabolic inequalities

$$|u_t - \mathbf{B}_0 u| \le M \textstyle\sum_{0 \le k < 2m} |D^k u| \quad \text{in } Q_1, \tag{7.3}$$

where $M > 0$ is a constant and $D^k u$ is the vector $\{D^\beta u,\ |\beta| = k\}$. These inequalities include the parabolic PDE (1.1) with constant coefficient $a_\beta = A_\beta$ for $|\beta| = 2m$ and arbitrary uniformly bounded coefficients $|a_\beta| \le M$ in the lower-order operators with $|\beta| < 2m$. We then arrive at a similar result.

Theorem 7.2. *Let $u(\cdot, t) \in \tilde{H}^{2m}_{\rho^*}$ be a solution of* (7.3), *and* (7.1) *hold. Then $u = 0$.*

Proof. One can see that after Sturmian scaling (1.4), (1.5), the function $u(y, \tau)$ can be treated as a solution of the PDE (1.7) where the perturbation $\mathbf{C}(\tau)$ is uniformly exponentially small as $\tau \to \infty$, and the above conclusion applies. □

The present approach to multiple zero formations and the corresponding unique continuation theorems admits extensions to quasilinear uniformly parabolic PDEs

$$u_t = \textstyle\sum_{|\beta| \le 2m} a_\beta(x, t, u) D^\beta u \quad \text{in } Q_1, \tag{7.4}$$

with sufficiently smooth bounded coefficients $a_\beta(x, t, u)$ satisfying necessary hypotheses. Then $A_\beta = a_\beta(0, 0, 0)$ for $|\beta| = 2m$ and $\mathbf{B}_0$ is assumed to be uniformly elliptic.

8. Dimension of nodal sets

Without loss of generality, we formulate the result on the Hausdorff dimension of nodal sets for the solutions of parabolic inequalities (7.3).

Theorem 8.1. *Let, under given hypotheses, $u(\cdot, t) \in \tilde{H}^{2m}_{\rho^*}$, $u \not\equiv 0$, be a sufficiently smooth solution of* (7.3). *Then its nodal set* (1.18) *satisfies* (1.19).

Estimates like (1.19) are well known for the second-order elliptic and parabolic equations with the proof based on a general idea of the dimensional reduction argument in the geometric measure theory; see Section 2 in [17] and [14].

Proof. We follow the lines of the analysis given in [3], Sections 8 and 9, which can be applies to solutions of higher-order inequalities or equations (or other functions exhibiting suitable asymptotic scaling properties at any point (x_0, t_0)) provided that two crucial results are available:

(i) The result of Corollary 6.2 makes it possible to introduce a locally asymptotically self-similar pair $(\mathcal{F}, \mathcal{L})$ as in [3], p. 627, where $\mathcal{F}$ is a collection of sufficiently smooth solutions u and $\mathcal{L}[u] = \{(x, 0) \in \mathbf{R}^N \times \mathbf{R} :\ u(x, 0) = 0\}$. The only difference is that according to (6.36) we define the scaling map $g(y, s; \lambda, \alpha)$ as follows:

$$(g(y, s; \lambda, \alpha)u)(x, t) = \alpha u(y + \lambda x, s + \lambda^{2m} t).$$

(ii) The polynomial structure of the limit function W in (6.39) makes it possible to apply Theorem 8.5 in [3] and to complete the proof. □

Estimates on the parabolic dimension of various nodal sets obtained in [3] for $m = 1$ remain valid for higher-order differential parabolic operators.

Acknowledgement. The author would like to thank Yu.V. Egorov and S.I. Pohozaev for useful discussions.

REFERENCES

[1] S.B. Angenent, *The Morse-Smale property for a semi-linear parabolic equation,* J. Differ. Equat., **62** (1986), 427-442.

[2] M.S. Birman and M.Z. Solomjak, Spectral Theory of Self-Adjoint Operators in Hilbert Space, D. Reidel, Dordrecht/Tokyo, 1987.

[3] X.-Y. Chen, *A strong unique continuation theorem for parabolic equations*, Math. Ann., **311** (1998), 603-630.

[4] M. Chen, X.-Y. Chen, and J.K. Hale, *Structural stability for time-periodic one-dimensional parabolic equations*, J. Differ. Equat., **96** (1992), 355-418.

[5] Yu.V. Egorov, V.A. Galaktionov, V.A. Kondratiev, and S.I. Pohozaev, *Asymptotic behaviour of global solutions to higher-order semilinear parabolic equations in the supercritical range*, Comptes Rendus Acad. Sci. Paris, Série I, **335** (2002), 805-810 (full text in `http://www.maths.bath.ac.uk/MATHEMATICS/preprints.html`).

[6] S.D. Eidelman, Parabolic Systems, North-Holland Publ. Comp., Amsterdam/London, 1969.

[7] A. Friedman, Partial Differential Equations, Robert E. Krieger Publ. Comp., Malabar, 1983.

[8] I. Gohberg, S. Goldberg, and M.A. Kaashoek, Classes of Linear Operators, Vol. **1**, Operator Theory: Advances and Applications, Vol. **49**, Birkhäuser Verlag, Basel/Berlin, 1990.

[9] G.H. Hardy, *Note on a theorem of Hilbert*, Math. Z., **6** (1920), 314-317.

[10] H.P. Heinig, *Weighted norm inequalities for classes of operators*, Indiana Univ. Math. J., **33** (1984), 573-582.

[11] D. Henry, Geometric Theory of Semilinear Parabolic Equations, Lecture Notes in Math., Vol. **840**, Springer-Verlag, Berlin/Hong Kong, 1981.

[12] D.B. Henry, *Some infinite-dimensional Morse-Smale systems defined by parabolic partial differential equations*, J. Differ. Equat., **59** (1985), 165-205.

[13] A.N. Kolmogorov and S.V. Fomin, Elements of the Theory of Functions and Functional Analysis, Nauka, Moscow, 1976.

[14] F.-H. Lin, *Nodal sets of solutions of elliptic and parabolic equations*, Comm. Pure Appl. Math., **44** (1991), 287-308.

[15] V. Maz'ja, Sobolev Spaces, Springer-Verlag, Berlin/Tokyo, 1985.

[16] O.A. Oleinik and E.V. Radkevich, *Method of introducing of a parameter for evolution equations*, Russian Math. Surveys, **33** (1978), 7-84.

[17] L. Simon, Lectures on Geometric Measure Theory, Vol. **3**, Proc. Center for Mathematical Analysis, Austr. Nat. Univ., 1984.

[18] C. Sturm, *Mémoire sur une classe d'équations à différences partielles,* J. Math. Pures Appl., **1** (1836), 373-444.

DEPARTMENT OF MATHEMATICAL SCIENCES, UNIVERSITY OF BATH, BATH BA2 7AY, UK AND KELDYSH INSTITUTE OF APPLIED MATHEMATICS, MIUSSKAYA SQ. 4, 125047 MOSCOW, RUSSIA

E-mail address: `vag@maths.bath.ac.uk`